

Better Agnostic Clustering via Relaxed Tensor Norms

Pravesh K. Kothari

Jacob Steinhardt

November 21, 2017

We develop a new family of convex relaxations for k -means clustering based on sum-of-squares norms, a relaxation of the injective tensor norm that is efficiently computable using the Sum-of-Squares algorithm. We give an algorithm based on this relaxation that recovers a faithful approximation to the true means in the given data whenever the low-degree moments of the points in each cluster have bounded sum-of-squares norms.

We then prove a sharp upper bound on the sum-of-squares norms for moment tensors of any distribution that satisfies the *Poincaré inequality*. The Poincaré inequality is a central inequality in probability theory, and a large class of distributions satisfy it including Gaussians, product distributions, strongly log-concave distributions, and any sum or uniformly continuous transformation of such distributions.

As an immediate corollary, for any $\gamma > 0$, we obtain an efficient algorithm for learning the means of a mixture of k arbitrary Poincaré distributions in \mathbb{R}^d in time $d^{O(1/\gamma)}$ so long as the means have separation $\Omega(k^\gamma)$. This in particular yields an algorithm for learning Gaussian mixtures with separation $\Omega(k^\gamma)$, thus partially resolving an open problem of Regev and Vijayaraghavan [RV17].

Our algorithm works even in the robust setting where an ε fraction of arbitrary outliers are added to the data, as long as the fraction of outliers is smaller than the smallest cluster. We therefore obtain results in the strong agnostic setting where, in addition to not knowing the distribution family, the data itself may be arbitrarily corrupted.

1 Introduction

Progress on many fundamental unsupervised learning tasks has required circumventing a plethora of intractability results by coming up with natural restrictions on input instances that preserve some essential character of the problem. For example, while k -means clustering is NP-hard in the worst-case [MNV09], there is an influential line of work providing spectral algorithms for clustering mixture models satisfying appropriate assumptions [AM05, KK10, AS12]. On the flip side, we run the risk of developing algorithmic strategies that exploit strong assumptions in a way that makes them brittle. We are thus forced to walk the tight rope of avoiding computational intractability without “overfitting” our algorithmic strategies to idealized assumptions on input data.

Consider, for example, the problem of clustering data into k groups. On the one hand, a line of work leading to [AS12] shows that a variant of spectral clustering can recover the underlying clustering so long as each cluster has bounded covariance around its center and the cluster centers are separated by at least $\Omega(\sqrt{k})$. Known results can improve on this bound to require a separation of $\Omega(k^{1/4})$ if the cluster distributions are assumed to be isotropic and log-concave [VW02]. If the cluster means are in general position, other lines of work yields results for Gaussians [KMV10, MV10, BS10, HK13, BCV14, GVX14, BCMV14, ABG⁺14, GHK15] or for distributions satisfying independence assumptions [HKZ09, AGH⁺13]. However, the assumptions often play a crucial role in the algorithm. For example, the famous method of moments that yields a result for learning mixtures of Gaussians in general position uses the specific algebraic structure of the moment tensor of Gaussian distributions. Such techniques are unlikely to work for more general classes of distributions.

As another example, consider the robust mean estimation problem which has been actively investigated recently. Lai et. al. [LRV16] and later improvements [DKK⁺17, SCV18] show how to estimate the mean of an unknown distribution (with bounded second moments) where an ε fraction of points are adversarially corrupted, obtaining additive error $O(\sqrt{\varepsilon})$. On the other hand, Diakonikolas et. al. [DKK⁺16] showed how to estimate the mean of a Gaussian or product distribution with nearly optimal additive error $\tilde{O}(\varepsilon)$. However, their algorithm again makes strong use of the known algebraic structure of the moments of these distributions.

Further scrutiny reveals that the two examples of clustering and robust mean estimation suffer from a “second-moment” barrier. For both problems, the most general results algorithmically exploit only some boundedness condition on the second moments of the data, while the strongest results use *exact* information about higher moments (e.g. by assuming Gaussianity) and are thus brittle. This leads to the key conceptual driving force of the present work:

Can we algorithmically exploit boundedness information about a limited number of low-degree moments?

As the above examples illustrate, this is a natural way to formulate the “in-between” case between the two well-explored extremes. From an algorithmic perspective, this question forces us to develop techniques that can utilize information about higher moments of data for problems such as clustering and mean estimation. For these problems, we can more concretely ask:

Can we beat the second-moment barrier in the agnostic setting for clustering and robust mean estimation?

The term *agnostic* here refers to the fact that we want our algorithm to work for as wide a class of distributions as possible, and in particular to avoid making parametric assumptions (such as Gaussianity) about the underlying distribution.

The main goal of this work is to present a principled way to utilize higher moment information in input data and break the second moment barrier for both clustering and robust mean estimation. A key primitive in our approach is algorithmic certificates upper bounding the injective norms of *moment tensors* of data.

Given input points, consider the injective tensor norm of their moments that generalizes the spectral norm of a matrix:

$$\sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \langle x_i, v \rangle^{2t}. \quad (1.1)$$

For $t > 1$, bounds on the injective norm of the moment tensor present a natural way to utilize higher moment information in the given data, which suggests an avenue for algorithm design. Indeed, one of our contributions (Theorem 1.2) is a generalization of spectral norm clustering that uses estimates of injective norms of moment tensors to go beyond the second moment barrier.

Unfortunately for us, estimating injective norms (unlike the spectral norm) is intractable. While it is likely easier than computing injective norms for arbitrary tensors, it turns out that approximately computing injective norms for moment tensors is equivalent to the well-studied problem of approximating the $2 \rightarrow q$ norm which is known to be small-set-expansion hard [BBH⁺12b]. The best known algorithms for approximating $2 \rightarrow q$ norm achieve a multiplicative approximation ratio of $d^{\Theta(q)}$ in d dimensions, and while known hardness results [BBH⁺12b] only rule out some fixed constant factor algorithms for this problem, it seems likely that there is no polynomial time algorithm for $2 \rightarrow q$ norm that achieves *any* dimension-independent approximation ratio.

An average-case variant of approximating injective norms of moment tensors has been studied to some extent due to its relationship to the small-set-expansion problem. The sum-of-squares hierarchy of semi-definite programming relaxations turns out to be a natural candidate algorithm in this setting and is known to exactly compute the injective norm in specialized settings such as that of the Gaussian distribution. On the other hand, the most general such results [BBH⁺12b, BKS15] imply useful bounds only for settings similar to product distributions.

One of the key technical contributions of this work is to go beyond product distributions for estimating injective norms. Specifically, we show (Theorem 1.1) that Sum-of-Squares gives a polynomial time procedure to show a dimension-free upper bound on the injective norms of (large enough i.i.d. samples from) arbitrary distributions that satisfy a *Poincaré inequality*. This is a much more satisfying state of affairs as it immediately captures all strongly log-concave distributions, including correlated Gaussians. Further, the Poincaré inequality is robust—i.e., it continues to hold under uniformly continuous transformations of the underlying space, as well as bounded re-weightings of the probability density.

Without further ado, we define Poincaré distributions: A distribution p on \mathbb{R}^d is said to be σ -Poincaré if for all differentiable functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ we have

$$\text{Var}_{x \sim p}[f(x)] \leq \sigma^2 \mathbb{E}_{x \sim p}[\|\nabla f(x)\|_2^2]. \quad (1.2)$$

This is a type of isoperimetric inequality on the distribution x and implies concentration of measure. In Section 3 we discuss in more detail various examples of distributions that satisfy (1.2), as well as properties of such distributions. Poincaré inequalities and distributions are intensely studied in probability theory; indeed, we rely on one such powerful result of Adamczak and Wolff [AW15] for establishing a sharp bound on the sum-of-squares algorithm’s estimate of the injective norm of an i.i.d. sample from a Poincaré distribution.

We then confirm the intuitive claim that understanding injective norms of moment tensors can give us an algorithmic tool to beat the second moment barrier, by combining our result on certification of Poincaré distributions with our algorithm for clustering under such certificates. Specifically, we show that for any $\gamma > 0$, given a balanced mixture of k Poincaré distributions with means separated by $\Omega(k^\gamma)$, we can successfully cluster n samples from this mixture in $n^{O(1/\gamma)}$ time (by using $O(1/\gamma)$ levels of the sum-of-squares hierarchy). Similarly, given samples from a Poincaré distribution with an ε fraction of adversarial corruptions, we can estimate its mean up to an error of $O(\varepsilon^{1-\gamma})$ in $n^{O(1/\gamma)}$ time. In fact, we will see below that we get both at once: a robust clustering algorithm that can learn well-separated mixtures even in the presence of arbitrary outliers.

To our knowledge such a result was not previously known even in the second-moment case ([CSV17] and [SCV18] study this setting but only obtain results in the *list-decodable* learning model). Our result only relies on the SOS-certifiability of the moment tensor, and holds for any deterministic point set for which such a sum-of-squares certificate exists.

Despite their generality, our results are strong enough to yield new bounds even in very specific settings such as learning balanced mixtures of k spherical Gaussians with separation $\Omega(k^\gamma)$. Our algorithm allows recovering the true means in $n^{O(1/\gamma)}$ time and partially resolves an open problem posed in the recent work of [RV17].

Certifying injective norms of moment tensors appears to be a useful primitive and could help enable further applications of the sum of squares method in machine learning. Indeed, [KS17] studies the problem of robust estimation of higher moments of distributions that satisfy a bounded-moment condition closely related to approximating injective norms. Their relaxation and the analysis are significantly different from the present work; nevertheless, our result for Poincaré distributions immediately implies that the robust moment estimation algorithm of [KS17] succeeds for a large class of Poincaré distributions.

1.1 Main Results and Applications

Our first main result regards efficient upper bounds on the injective norm of the moment tensor of any Poincaré distribution. Let $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ be n i.i.d. samples from a Poincaré distribution with mean μ , and let $M_{2t} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^{\otimes 2t}$ be the empirical estimate of the $2t$ th moment tensor. We are interested in upper-bounding the injective norm (1.1), which can be equivalently expressed in terms of the moment tensor as

$$\sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \langle x_i - \mu, v \rangle^{2t} = \sup_{\|v\|_2 \leq 1} \langle M_{2t}, v^{\otimes 2t} \rangle. \quad (1.3)$$

Standard results (see Fact 3.5) yield dimension-free upper bounds on (1.3) for all Poincaré distributions. Our first result is a “sum-of-squares proof” of this fact giving an *efficient* method to certify dimension-free upper bounds on (1.3) for samples from any Poincaré distribution.

Specifically, let the sum of squares norm of M_{2t} , denoted by $\|M_{2t}\|_{\text{sos}_{2t}}$, be the degree- $2t$ sum-of-squares relaxation of (1.3) (we discuss such norms and the sum-of-squares method in more detail in Section 2; for now the important fact is that $\|M_{2t}\|_{\text{sos}_{2t}}$ can be computed in time $(nd)^{O(t)}$). We show that for a large enough sample from a distribution that satisfies the Poincaré inequality, the sum-of-squares norm of the moment tensor is upper bounded by a dimension-free constant.

Theorem 1.1. *Let p be a σ -Poincaré distribution over \mathbb{R}^d with mean μ . Let $x_1, \dots, x_n \sim p$ with $n \geq (2d \log(dt/\delta))^t$. Then, for some constant C_t (depending only on t) with probability at least $1 - \delta$ we have $\|M_{2t}\|_{\text{sos}_{2t}} \leq C_t \sigma$, where $M_{2t} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^{\otimes 2t}$.*

As noted above, previous sum-of-squares bounds worked for specialized cases such as product distributions. Theorem 1.1 is key to our applications that crucially rely on 1) going beyond product distributions and 2) using sos_{2t} norms as a proxy for injective norms for higher moment tensors.

Outlier-Robust Agnostic Clustering. Our second main result is an efficient algorithm for *outlier-robust agnostic clustering* whenever the “ground-truth” clusters have moment tensors with bounded sum-of-squares norms.

Concretely, the input is data points x_1, \dots, x_n of n points in \mathbb{R}^d , a $(1 - \varepsilon)$ fraction of which admit a (unknown) partition into sets I_1, \dots, I_k each having bounded sum-of-squares norm around their corresponding means μ_1, \dots, μ_k . The remaining ε fraction can be arbitrary outliers. Observe that in this setting, we do not make any explicit distributional assumptions.

We will be able to obtain strong estimation guarantees in this setting so long as the clusters are *well-separated* and the fraction ε of outliers is not more than $\alpha/8$, where α is the fraction of points in the smallest cluster. We define the separation as $\Delta = \min_{i \neq j} \|\mu_i - \mu_j\|_2$. A lower bound on Δ is information theoretically necessary even in the special case of learning mixtures of identity-covariance gaussians without any outliers (see [RV17]).

Theorem 1.2. *Suppose points $x_1, \dots, x_n \in \mathbb{R}^d$ can be partitioned into sets I_1, \dots, I_k and *out*, where the I_j are the clusters and *out* is a set of outliers of size εn . Suppose I_j has size $\alpha_j n$ and mean μ_j , and that its $2t$ th moment $M_{2t}(I_j)$ satisfies $\|M_{2t}(I_j)\|_{\text{sos}_{2t}} \leq B$. Also suppose that $\varepsilon \leq \alpha/8$ for $\alpha = \min_{j=1}^k \alpha_j$.*

Finally, suppose the separation $\Delta \geq C_{\text{sep}} \cdot B/\alpha^{1/t}$, with $C_{\text{sep}} \geq C_0$ (for a universal constant C_0). Then there is an algorithm running in time $(nd)^{O(t)}$ and outputting means $\hat{\mu}_1, \dots, \hat{\mu}_k$ such that $\|\hat{\mu}_j - \mu_j\|_2 \leq O(B(\varepsilon/\alpha + C_{\text{sep}}^{-2t})^{1-1/2t})$ for all j .

The parameter B specifies a bound on the variation in each cluster. The separation condition says that the distance between cluster means must be slightly larger (by a $\alpha^{-1/t}$ factor) than this variation. The error in recovering the cluster means depends on two terms—the fraction of outliers ε , and the separation C_{sep} .

To understand the guarantees of the theorem, let’s start with the case where $\varepsilon = 0$ (no outliers) and $\alpha = 1/k$ (all clusters have the same size). In this case, the separation requirement between the clusters is $B \cdot k^{1/t}$ where B is the bound on the moment tensor of order $2t$. The theorem guarantees a recovery of the means up to an error in Euclidean norm of $O(B)$. By taking t larger (and spending the correspondingly larger running time), our clustering algorithm works with separation k^γ for any constant γ . This is the first result that goes beyond the separation requirement of $k^{1/2}$ in the *agnostic* clustering setting—i.e., without making distributional assumptions on the clusters.

It is important to note that even in 1 dimension, it is information theoretically impossible to recover cluster means to an error $\ll B$ when relying only on $2t$ th moment bounds. A simple example to illustrate this is obtained by taking a mixture of two distributions on the real line with bounded $2t$ th moments but small overlap in the tails. In this case, it is impossible to correctly classify the points that come from the overlapping part. Thus, a fraction of points in the tail always end up misclassified, shifting the true means. The recovery error of our algorithm does indeed drop as the separation (controlled by C_{sep}) between the true means increases (making the overlapping parts of the tail smaller). We note that for the specific case of spherical gaussians, we can exploit their parametric structure to get arbitrarily accurate estimates even for fixed separation; see Corollary 1.4.

Next, let's consider $\varepsilon \neq 0$. In this case, if $\varepsilon \ll \alpha$, we recover the means up to an error of $O(B)$ again (for $C_{\text{sep}} \geq C_0$). It is intuitive that the recovery error for the means should grow with the number of outliers, and the condition $\varepsilon \leq \alpha/8$ is necessary, as if $\varepsilon \geq \alpha$ then the outliers could form an entirely new cluster making recovery of the means information-theoretically impossible.

We also note that in the degenerate case where $k = 1$ (a single cluster), Theorem 1.2 yields results for robust mean estimation of a set of points corrupted by an ε fraction of outliers. In this case we are able to estimate the mean to error $\varepsilon^{\frac{2t-1}{2t}}$; when $t = 1$ this is $\sqrt{\varepsilon}$, which matches the error obtained by methods based on second moments [LRV16, DKK⁺17, SCV18]. For $t = 2$ we get error $\varepsilon^{3/4}$, for $t = 3$ we get error $\varepsilon^{5/6}$, and so on, approaching an error of ε as $t \rightarrow \infty$. In particular, this pleasingly approaches the rate $\tilde{O}(\varepsilon)$ obtained by much more bespoke methods that rely strongly on specific distributional assumptions [LRV16, DKK⁺16].

Note that we could not hope to do better than $\varepsilon^{\frac{2t-1}{2t}}$, as that is the information-theoretically optimal error for distributions with bounded $2t$ th moments (even in one dimension), and degree- $2t$ SOS only “knows about” moments up to $2t$.

Finally, we can obtain results even for clusters that are not well-separated, and for fractions of outliers that could exceed α . In this case we no longer output exactly k means, and must instead consider the list-decodable model [BBV08, CSV17], where we output a list of $O(1/\alpha)$ means of which the true means are a sublist. We defer the statement of this result to Theorem 5.5 in Section 5.

Applications. Putting together Theorem 1.1 and Theorem 1.2 immediately yields corollaries for learning mixtures of Poincaré distributions, and in particular mixtures of Gaussians.

Corollary 1.3 (Disentangling Mixtures of Arbitrary Poincaré Distributions). *Suppose that we are given a dataset of n points x_1, \dots, x_n , such that at least $(1 - \varepsilon)n$ points are drawn from a mixture $\alpha_1 p_1 + \dots + \alpha_k p_k$ of k distributions, where p_j is σ -Poincaré with mean μ_j (the remaining εn points may be arbitrary). Let $\alpha = \min_{j=1}^k \alpha_j$. Also suppose that the separation Δ is at least $C_{\text{sep}} \cdot C_t \sigma / \alpha^{1/t}$, for some constant C_t depending only on t and some $C_{\text{sep}} \geq 1$.*

Then, assuming that $\varepsilon \leq \frac{\alpha}{10}$, for some $n = O((2d \log(tk d/\delta))^t / \alpha + d \log(k/\delta) / \alpha \varepsilon^2)$, there is an algorithm running in $n^{O(t)}$ time which with probability $1 - \delta$ outputs candidate means $\hat{\mu}_1, \dots, \hat{\mu}_k$ such that $\|\hat{\mu}_j - \mu_j\|_2 \leq C'_t \sigma (\varepsilon/\alpha + C_{\text{sep}}^{-2t})^{\frac{2t-1}{2t}}$ for all j (where C'_t is a different universal constant).

The $1/\alpha$ factor in the sample complexity is so that we have enough samples from every single cluster for Theorem 1.1 to hold. The extra term of $d \log(k/\delta) / \varepsilon^2$ in the sample complexity is so that the empirical means of each cluster concentrate to the true means.

Corollary 1.3 is one of the strongest results on learning mixtures that one could hope for. If the mixture weights α are all at least $1/\text{poly}(k)$, then Corollary 1.3 implies that we can cluster the points as long as the separation $\Delta = \Omega(k^\gamma)$ for any $\gamma > 0$. Even for spherical Gaussians the best previously known algorithms required separation $\Omega(k^{1/4})$. On the other hand, Corollary 1.3 applies to a large family of distributions including arbitrary strongly log-concave distributions. Moreover, while the Poincaré inequality does not directly hold for discrete distributions, Fact 3.3 in Section 3 implies that a large class of discrete distributions, including product distributions over bounded domains, will satisfy the Poincaré inequality after adding zero-mean Gaussian noise. Corollary 1.3 then yields a clustering algorithm for these distributions, as well.

For mixtures of Gaussians in particular, we can do better, and in fact achieve vanishing error independent of the separation:

Corollary 1.4 (Learning Mixtures of Gaussians). *Suppose that $x_1, \dots, x_n \in \mathbb{R}^d$ are drawn from a mixture of k Gaussians: $p = \sum_{j=1}^k \alpha_j \mathcal{N}(\mu_j, I)$, where $\alpha_j \geq 1/\text{poly}(k)$ for all j . Then for any $\gamma > 0$, there is a separation $\Delta_0 = O(k^\gamma)$ such that given $n \geq \text{poly}(d^{1/\gamma}, k, 1/\varepsilon) \log(k/\delta)$ samples from p , if the separation $\Delta \geq \Delta_0$, then with probability $1 - \delta$ we obtain estimates $\hat{\mu}_1, \dots, \hat{\mu}_k$ with $\|\hat{\mu}_j - \mu_j\|_2 \leq \varepsilon$ for all j .*

Remark 1.5. This partially resolves an open question of [RV17], who ask whether it is possible to efficiently learn mixtures of Gaussians with separation $\sqrt{\log k}$.

The error now goes to 0 as $n \rightarrow \infty$, which is not true in the more general Corollary 1.3. This requires invoking Theorem IV.1 of [RV17], which, given a sufficiently good initial estimate of the means of a mixture of Gaussians, shows how to get an arbitrarily accurate estimate. As discussed before, such a result is specific to Gaussians and in particular is information-theoretically impossible for mixtures of general Poincaré distributions.

1.2 Proof Sketch and Technical Contributions

We next sketch the proofs of our two main theorems (Theorem 1.1 and Theorem 1.2) while indicating which parts involve new technical ideas.

1.2.1 Sketch of Theorem 1.1

For simplicity, we will only focus on SOS-certifiability in the infinite-data limit, i.e. on showing that SOS can certify an upper bound $\mathbb{E}_{x \sim p}[\langle x - \mu, v \rangle^{2t}] \leq C_t \sigma^{2t} \|v\|_2^{2t}$. (In Section 4.2 we will show that finite-sample concentration follows due to the matrix Rosenthal inequality [MJC⁺14].)

We make extensive use of a result of [AW15]; it is a very general result on bounding non-Lipschitz functions of Poincaré distributions, but in our context the important consequence is the following:

If $f(x)$ is a degree- t polynomial such that $\mathbb{E}_p[\nabla^j f(x)] = 0$ for $j = 0, \dots, t-1$, then $\mathbb{E}_p[f(x)^2] \leq C_t \sigma^{2t} \|\nabla^t f(x)\|_F^2$ for a constant C_t , assuming p is σ -Poincaré. (Note that $\nabla^t f(x)$ is a constant since f is degree- t .)

Here $\|A\|_F^2$ denotes the Frobenius norm of the tensor A , i.e. the ℓ_2 -norm of A if it were flattened into a d^t -element vector.

We can already see why this sort of bound might be useful for $t = 1$. Then if we let $f_v(x) = \langle x - \mu, v \rangle$, we have $\mathbb{E}[f_v(x)] = 0$ and hence $\mathbb{E}_p[\langle x - \mu, v \rangle^2] \leq C_1 \sigma^2 \|v\|_2^2$. This exactly says that p has bounded covariance.

More interesting is the case $t = 2$. Here we will let $f_A(x) = \langle (x - \mu)(x - \mu)^\top - \Sigma, A \rangle$, where μ is the mean and Σ is the covariance of p . It is easy to see that both $\mathbb{E}[f_A(x)] = 0$ and $\mathbb{E}[\nabla f_A(x)] = 0$. Therefore, we have $\mathbb{E}[\langle (x - \mu)(x - \mu)^\top - \Sigma, A \rangle^2] \leq C_2 \sigma^4 \|A\|_F^2$.

Why is this bound useful? It says that if we unroll $(x - \mu)(x - \mu)^\top - \Sigma$ to a d^2 -dimensional vector, then this vector has bounded covariance (since if we project along any direction A with $\|A\|_F = 1$, the variance is at most $C_2 \sigma^4$). This is useful because it turns out sum-of-squares “knows about” such covariance bounds; indeed, this type of covariance bound is exactly the property used in [BBH⁺12a] to certify 4th moment tensors over the hypercube. In our case it yields a sum-of-squares proof that $\mathbb{E}[\langle (x - \mu)^{\otimes 4} - \Sigma^{\otimes 2}, v^{\otimes 4} \rangle] \leq_{\text{sos}} C_2 \sigma^4 \|v\|_2^4$, which can then be used to bound the 4th moment $\mathbb{E}[\langle x - \mu, v \rangle^4] = \mathbb{E}[\langle (x - \mu)^{\otimes 4}, v^{\otimes 4} \rangle]$.

Motivated by this, it is natural to try the same idea of “subtracting off the mean and squaring” with $t = 4$. Perhaps we could define $f_A(x) = \langle ((x - \mu)^{\otimes 2} - \Sigma)^{\otimes 2} - \mathbb{E}[\langle (x - \mu)^{\otimes 2} - \Sigma \rangle^2], A \rangle$?

Alas, this does not work—while there is a suitable polynomial $f_A(x)$ for $t = 4$ that yields sum-of-squares bounds, it is somewhat more subtle. For simplicity we will write the polynomial for $t = 3$. It is the following: $f_A(x) = \langle (x - \mu)^{\otimes 3} - 3(x - \mu) \otimes \Sigma - M_3, A \rangle$, where $M_3 = \mathbb{E}[(x - \mu)^{\otimes 3}]$ is the third-moment tensor of p . By checking that $\mathbb{E}[f_A(x)] = \mathbb{E}[\nabla f_A(x)] = \mathbb{E}[\nabla^2 f_A(x)] = 0$, we obtain that the tensor $F_3(x) = (x - \mu)^{\otimes 3} - 3(x - \mu) \otimes \Sigma - M_3$, when unrolled to a d^3 -dimensional vector, has bounded covariance, which means that sum-of-squares knows that $\mathbb{E}[\langle F_3(x)^{\otimes 2}, v^{\otimes 6} \rangle]$ is bounded for all $\|v\|_2 \leq 1$.

However, this is not quite what we want—we wanted to show that $\mathbb{E}[\langle (x - \mu)^{\otimes 6}, v^{\otimes 6} \rangle]$ is bounded. Fortunately, the leading term of $F_3(x)^{\otimes 2}$ is indeed $(x - \mu)^{\otimes 6}$, and all the remaining terms are lower-order. So, we can subtract off $F_3(x)$ and recursively bound all of the lower-order terms to get a sum-of-squares bound on $\mathbb{E}[\langle (x - \mu)^{\otimes 6}, v^{\otimes 6} \rangle]$. The case of general t follows similarly, by carefully constructing a tensor $F_t(x)$ whose first $t - 1$ derivatives are all zero in expectation.

There are a couple contributions here beyond what was known before. The first is identifying appropriate tensors $F_t(x)$ whose covariances are actually bounded so that sum-of-squares can make use of them. For $t = 1, 2$ (the cases that had previously been studied) the appropriate tensor is in some sense the “obvious” one $(x - \mu)^{\otimes 2} - \Sigma$, but even for $t = 3$ we end up with the fairly non-obvious tensor $(x - \mu)^{\otimes 3} - 3(x - \mu) \otimes \Sigma - M_3$. (For $t = 4$ it is $(x - \mu)^{\otimes 4} - 6(x - \mu)^{\otimes 2} \otimes \Sigma - 4(x - \mu) \otimes M_3 - M_4 + 6\Sigma \otimes \Sigma$.) While these tensors may seem mysterious a priori, they are actually the unique tensor polynomials with leading term $x^{\otimes t}$ such that all derivatives of order $j < t$ have mean zero. Even beyond Poincaré distributions, these seem like useful building blocks for sum-of-squares proofs.

The second contribution is making the connection between Poincaré distributions and the above polynomial inequalities. The well known work of Latała [Lat06] establishes non-trivial estimates of upper bounds on the moments of polynomials of Gaussians, of which the inequalities used here are a special case. [AW15] show that these inequalities also hold for Poincaré distributions. However, it is not a priori obvious that these inequalities should lead to sum-of-squares proofs, and it requires a careful invocation of the general inequalities to get the desired results in the present setting.

1.2.2 Sketch of Theorem 1.2

We next establish our result on robust clustering. In fact we will establish a robust mean estimation result which will lead to the clustering result—specifically, we will show that if a set of points x_1, \dots, x_n contains a subset $\{x_i\}_{i \in I}$ of size αn that is SOS-certifiable, then the mean (of the points in I) can be estimated regardless of the remaining points. There are two parts: if $\alpha \approx 1$ we want to show error going to 0 as $\alpha \rightarrow 1$, while if $\alpha \ll 1$ we want to show error that does not grow too fast as $\alpha \rightarrow 0$. In the latter case we will output $O(1/\alpha)$ candidates for the mean and show that at least one of them is close to the true mean (think of these candidates as accounting for $O(1/\alpha)$ possible clusters in the data). We will later prune down to exactly k means for well-separated clusters.

For $t = 1$ (which corresponds to bounded covariance), the $\alpha \rightarrow 0$ case is studied in [CSV17]. A careful analysis of the proof there reveals that all of the relevant inequalities are sum-of-squares inequalities, so there is a sum-of-squares generalization of the algorithm in [CSV17] that should give bounds for SOS-certifiable distributions. While this would likely lead to some robust clustering result, we note the bounds we achieve here are stronger than those in [CSV17], as [CSV17] do not achieve tight results when the clusters are well-separated. Moreover, the proof in [CSV17] is complex and would be somewhat tedious to extend in full to the sum-of-squares setting.

We combine and simplify ideas from both [CSV17] and [SCV18] to obtain a relatively clean

algorithm. In fact, we will see that a certain mysterious constraint appearing in [CSV17] is actually the natural constraint from a sum-of-squares perspective.

Our algorithm is based on the following optimization. Given points x_1, \dots, x_n , we will try to find points w_1, \dots, w_n such that $\frac{1}{n} \sum_{i=1}^n \tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - w_i, v \rangle^{2t}]$ is small for all pseudodistributions ξ over the sphere. This is natural because we know that for the good points x_i and the true mean μ , $\langle x_i - \mu, v \rangle^{2t}$ is small (by the SOS-certifiability assumption). However, without further constraints this is not a very good idea because the trivial optimum is to set $w_i = x_i$. We would somehow like to ensure that the w_i cannot overfit too much to the x_i ; it turns out that the natural way to measure this degree of overfitting is via the quantity $\sum_{i \in I} \langle w_i - \mu, w_i \rangle^{2t}$.

Of course, this quantity is not known because we do not know μ . But we *do* know that $\sum_{i \in I} \tilde{\mathbb{E}}_{\xi(v)}[\langle w_i - \mu, v \rangle^{2t}]$ is small for all pseudodistributions (because the corresponding quantity is small for $x_i - \mu$ and $w_i - x_i$, and hence also for $w_i - \mu = (w_i - x_i) + (x_i - \mu)$ by Minkowski's inequality). Therefore, we impose the following constraint: *whenever z_1, \dots, z_n are such that $\sum_{i=1}^n \tilde{\mathbb{E}}_{\xi}[\langle z_i, v \rangle^{2t}] \leq 1$ for all ξ , it is also the case that $\sum_{i=1}^n \langle z_i, w_i \rangle^{2t}$ is small.* This constraint is not efficiently imposable, but it does have a simple sum-of-squares relaxation. Namely, we require that $\sum_{i=1}^n \langle Z_i, w_i^{\otimes 2t} \rangle$ is small whenever Z_1, \dots, Z_n are pseudomoment tensors satisfying $\sum_{i=1}^n Z_i \preceq_{\text{SOS}} I$.

Together, this leads to seeking w_1, \dots, w_n such that

$$\sum_{i=1}^n \tilde{\mathbb{E}}_{\xi}[\langle x_i - w_i, v \rangle^{2t}] \text{ is small for all } \xi, \text{ and } \sum_{i=1}^n \langle Z_i, w_i^{\otimes 2t} \rangle \text{ is small whenever } \sum_i Z_i \preceq_{\text{SOS}} I. \quad (1.4)$$

If we succeed in this, we can show that we end up with a good estimate of the mean (more specifically, the w_i are clustered into a small number of clusters, such that one of them is centered near μ). The above is a convex program, and thus, if this is impossible, by duality there must exist specific ξ and Z_1, \dots, Z_n such that the above quantities cannot be small for *any* w_1, \dots, w_n . But for fixed ξ and $Z_{1:n}$, the different w_i are independent of each other, and in particular it should be possible to make both sums small at least for the terms coming from the good set I . This gives us a way of performing *outlier removal*: look for terms where $\min_w \tilde{\mathbb{E}}_{\xi}[\langle x_i - w, v \rangle^{2t}]$ or $\min_w \langle Z_i, w \rangle$ is large, and remove those from the set of points. We can show that after a finite number of iterations this will have successfully removed many outliers and few good points, so that eventually we must succeed in making both sums small and thus get a successful clustering.

Up to this point the proof structure is similar to [SCV18]; the main innovation is the constraint involving the z_i , which bounds the degree of overfitting. In fact, when $t = 1$ this constraint is the dual form of one appearing in [CSV17], which asks that $w_i^{\otimes 2} \preceq Y$ for all i , for some matrix Y of small trace. In [CSV17], the matrix Y couples all of the variables, which complicates the analysis. In the form given here, we avoid the coupling and also see why the constraint is the natural one for controlling overfitting.

To finish the proof, it is also necessary to iteratively re-cluster the w_i and re-run the algorithm on each cluster. This is due to issues where we might have, say, 3 clusters, where the first two are relatively close together but very far from the third one. In this case our algorithm would resolve the third cluster from the first two, but needs to be run a second time to then resolve the first two clusters from each other.

[CSV17] also use this re-clustering idea, but their re-clustering algorithm makes use of a sophisticated metric embedding technique and is relatively complex. Here we avoid this complexity by making use of *resilient sets*, an idea introduced in [SCV18]. A resilient set is a set such that all large subsets have mean close to the mean of the original set; it can be shown that any set with

bounded moment tensor is resilient, and by finding such resilient sets we can robustly cluster in a much more direct manner than before. In particular, in the well-separated case we show that after enough rounds of re-clustering, every resilient set has almost all of its points coming from a single cluster, leading to substantially improved error bounds in that case.

1.3 Open Problems

In this work, we showed that sum-of-squares can certify moment tensors for distributions satisfying the Poincaré inequality. While this class of distributions is fairly broad, one could hope to establish sum-of-squares bounds for even broader families. Indeed, one canonical family is the class of *sub-Gaussian* distributions. Is it the case that sum-of-squares certifies moment tensors for all sub-Gaussian distributions? Conversely, are there sub-Gaussian distributions that sum-of-squares cannot certify? Even for 4th moments, this is unknown:

Open Question 1.6. Let p be a σ -sub-Gaussian distribution and let $M_4(p)$ denote its fourth moment tensor. Is it always the case that $\|M_4(p)\|_{\text{sos}_{2t}} \leq O(1) \cdot \sigma$ for some constant t ?

In another direction, the only property we required from Poincaré distributions is Adamczak and Wolff’s result [AW15] bounding the variance of polynomials whose derivatives all have mean 0. Adamczak and Wolff show that this property also holds for other distributions, such as sub-Gaussian product distributions. One might expect additional distributions to satisfy these inequalities as well, in which case our present results would apply unchanged.

Open Question 1.7. Say that a distribution p satisfies the (t, σ) -moment property if, whenever $f(x)$ is a degree- t polynomial with $\mathbb{E}[f(x)] = \mathbb{E}[\nabla f(x)] = \dots = \mathbb{E}[\nabla^{t-1} f(x)] = 0$, we have $\mathbb{E}[f(x)^2] \leq \sigma^{2t} \|\nabla^t f(x)\|_p^2$. Which distributions satisfy the (t, σ) -moment property?

Finally, the present results all regard certifying moment tensors in the ℓ_2 -norm, i.e., on upper bounding $\langle M_{2t}(p), v^{\otimes 2t} \rangle$ for all $\|v\|_2 \leq 1$. However, [SCV18] show that in some cases—such as discrete distribution learning—the ℓ_∞ -norm is more natural. To this end, define $\|M_{2t}\|_{\text{sos}_{2t}, \infty}$ to be the maximum of $\tilde{\mathbb{E}}_{\xi(v)}[\langle M_{2t}, v^{\otimes 2t} \rangle]$ over all pseudodistributions on the *hypercube*.

Open Question 1.8. For what distributions p is $\|M_{2t}\|_{\text{sos}_{2t}, \infty}$ small? Additionally, do bounds on $\|M_{2t}\|_{\text{sos}_{2t}, \infty}$ lead to better robust estimation and clustering in the ℓ_∞ -norm?

2 Preliminaries

In this section we set up notation and introduce a number of preliminaries regarding sum-of-squares algorithms.

Notation. We will use d to denote dimension, and n the number of samples in a dataset x_1, \dots, x_n . For clustering problems k will denote the number of clusters. ε will denote, depending on circumstance, either the desired estimation error or the fraction of adversarial corruptions for a robust estimation problem. δ will denote the probability of failure of an algorithm. γ will denote an exponent which we think of as going to zero, as in phrases like “ $O(k^\gamma)$ for any $\gamma > 0$ ”. For tensors, t will denote their order (or $2t$ if we want to emphasize the order is even). We let C denote a universal constant and C_t a universal constant depending on t (these constants may change in each place they are used).

Below we use Theorem (and Proposition, Lemma, etc.) for results that we prove in this paper, and Fact for results proved in other papers.

Tensors, Polynomials and Norms. A t th order tensor T on \mathbb{R}^d is a t -dimensional array of real numbers indexed by t -tuples on $[d]$. T is naturally associated with a homogenous degree t polynomial $T(v) = \langle T, v^{\otimes t} \rangle$. The *injective norm* of a tensor T is defined as $\sup_{\|v\|_2 \leq 1} T(v)$.

Given a distribution p on \mathbb{R}^d , the t th *moment tensor* of p is defined by $M_t(p) = \mathbb{E}_{x \sim p}[(x - \mu)^{\otimes t}]$, where μ is the mean of p . Observe that each entry of $M_t(p)$ is the expectation of some monomial of degree t with respect to p . For a finite set of points S , we let $\mathbb{E}_{x \sim S}$ denote expectation with respect to its empirical distribution. The moment tensor of a set of points is the moment tensor of its empirical distribution.

Given a matrix M , we let $\|M\|_{\text{op}}$ denote its operator norm (maximum singular value) and $\|M\|_F$ denotes its Frobenius norm (ℓ_2 -norm of its entries when flattened to a d^2 -dimensional vector). More generally, for a tensor T , we let $\|T\|_F$ denote the Frobenius norm (which is again the ℓ_2 -norm when the entries are flattened to a d^t -dimensional vector).

2.1 Sum-of-Squares Programs and Pseudodistributions

In this paper we are interested in approximating injective norms of moment tensors, i.e. of upper-bounding programs of the form

$$\text{maximize } \frac{1}{n} \sum_{i=1}^n \langle x_i, v \rangle^{2t} \quad (2.1)$$

$$\text{subject to } \|v\|_2 = 1. \quad (2.2)$$

This problem is hard to solve exactly, so we will instead consider the following *sum-of-squares relaxation* of (2.1):

$$\text{maximize } \frac{1}{n} \sum_{i=1}^n \tilde{\mathbb{E}}_{\xi(v)} [\langle x_i, v \rangle^{2t}] \quad (2.3)$$

$$\text{subject to } \tilde{\mathbb{E}}_{\xi(v)} [(\|v\|_2^2 - 1)p(v)] = 0 \text{ for all polynomials } p(v) \text{ of degree at most } 2t - 2, \quad (2.4)$$

$$\tilde{\mathbb{E}}_{\xi(v)} [q(v)^2] \geq 0 \text{ for all polynomials } q(v) \text{ of degree at most } t, \quad (2.5)$$

$$\tilde{\mathbb{E}}_{\xi(v)} [1] = 1. \quad (2.6)$$

Formally, $\tilde{\mathbb{E}}_{\xi(v)}$, which we refer to as a *pseudo-expectation*, is simply a linear functional on the space of polynomials in v of degree at most $2t$. The last two constraints say that ξ specifies a *pseudo-distribution*, meaning that it respects all properties of a regular probability distribution that can be specified with low-degree polynomials. Meanwhile, the first constraint is a relaxation of the constraint that $\|v\|_2 = 1$. If ξ satisfies (2.4-2.6), we say that ξ is a *pseudo-distribution on the sphere*.

The relaxation (2.3) is a special case of the well-studied family of sum-of-squares relaxations. It is well-known that these programs can be solved efficiently, i.e. in time $(nd)^{O(t)}$, due to the ability to represent them as semidefinite programs [Sho87, Par00, Nes00, Las01].

The key strategy for bounding the value of (2.3) is *sum-of-squares proofs*. We say that a polynomial inequality $p(v) \leq q(v)$ has a sum-of-squares proof if the polynomial $q(v) - p(v)$ can be written as

a sum of squares of polynomials. We write this as $p(v) \leq_{\text{sos}} q(v)$, or $p(v) \leq_{2t} q(v)$ if we want to emphasize that the proof only involves polynomials of degree at most $2t$.

For pseudo-distributions on the sphere, we will extend this notation and say that $p(v) \leq_{2t} q(v)$ if $q(v) - p(v) - r(v)(\|v\|_2^2 - 1)$ is a sum of squares for some polynomial $r(v)$ of degree at most $2t - 2$.

Now, let $p_0(v) = \frac{1}{n} \sum_{i=1}^n \langle x_i, v \rangle^{2t}$. Imagine that there is some sequence of sum-of-squares proofs $p_0(v) \leq_{2t} p_1(v) \leq_{2t} \dots \leq_{2t} p_m(v) = B^{2t} \cdot \|v\|_2^{2t}$. Then we know, by the constraints on ξ , that $\tilde{\mathbb{E}}_{\xi(v)}[p_0(v)] \leq B^{2t} \tilde{\mathbb{E}}_{\xi(v)}[\|v\|_2^{2t}] = B^{2t}$. Therefore, such a proof immediately implies that (2.3) has value at most B^{2t} .

Note that since the relation \leq_{2t} is transitive, this is equivalent to the condition $p_0(v) \leq_{2t} B^{2t} \|v\|_2^{2t}$. In this case, we call the set of points *SOS-certifiable*. More generally, for a distribution we have the following definition:

Definition 2.1. For a distribution p , we say that p is $(2t, B)$ -SOS-certifiable if $\langle M_{2t}(p), v^{\otimes 2t} \rangle \leq_{2t} B^{2t} \|v\|_2^{2t}$. We will alternately denote this by $\|M_{2t}(p)\|_{\text{sos}_{2t}} \leq B$. For a set of points x_1, \dots, x_n , we say it is $(2t, B)$ -SOS-certifiable if its empirical distribution is SOS-certifiable.

Note that $\langle M_{2t}(p), v^{\otimes 2t} \rangle = \mathbb{E}_{x \sim p}[\langle x, v \rangle^{2t}]$ so that this definition coincides with certifying (2.3).

2.2 Basic Sum-of-Squares Facts

We capture a few basic facts about sum-of-squares proofs that we will use later. First, sum-of-squares can certify all spectral norm bounds:

Fact 2.2 (Sum-of-Squares Proofs from Spectral Norm Bounds). *For any symmetric $d \times d$ matrix M , $\langle M, x^{\otimes 2} \rangle \leq_2 \|M\|_{\text{op}} \|x\|_2^2$.*

As a corollary (by applying the spectral norm bound to higher-order tensors), we also have the following:

Fact 2.3 (Spectral Norm Bounds for Tensors). *For a degree- t tensor function $F(x)$, suppose that $\mathbb{E}_{x \sim p}[\langle F(x), A \rangle^2] \leq \lambda \|A\|_F^2$ for all symmetric tensors A . Then, $\mathbb{E}_{x \sim p}[\langle F(x), v^{\otimes t} \rangle^2] \leq_{2t} \lambda \|v\|_2^{2t}$.*

Fact 2.3 is used crucially in the sequel. In fact, all of our sum-of-squares proofs will essentially involve invoking Fact 2.3 on appropriately chosen tensors $F(x)$.

Finally, the following basic inequality holds:

Fact 2.4. *If $0 \leq_t p(v) \leq_t p'(v)$ and $0 \leq_s q(v) \leq_s q'(v)$, then $p(v)q(v) \leq_{s+t} p'(v)q'(v)$.*

This is useful because it allows us to multiply sum-of-squares inequalities together.

3 Poincaré Distributions

In this section, we note some important properties of the class of all Poincaré distributions.

Definition 3.1 (Poincaré Distributions). A distribution p over \mathbb{R}^d is said to be σ -Poincaré if it satisfies the following Poincaré inequality with parameter σ : For all differentiable functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\text{Var}_{x \sim p}[f(x)] \leq \sigma^2 \mathbb{E}_{x \sim p}[\|\nabla f(x)\|_2^2]. \quad (3.1)$$

Note that no discrete distribution can satisfy (3.1). To see why, consider for instance the uniform distribution over $\{0, 1\}$. This cannot satisfy (3.1) for any σ , because for the function $f(x) = \max(0, \min(1, 3x - 2))$ we have $\text{Var}[f(x)] = \frac{1}{4}$ but $\mathbb{E}[\|\nabla f(x)\|_2^2] = 0$. More generally, (3.1) implies that there are no low-probability “valleys” separating two high-probability regions.

Next we give some examples of distributions satisfying (3.1). First, if $p = \mathcal{N}(\mu, \Sigma)$ is a normal distribution with mean μ and variance Σ , then p satisfies (3.1) with parameter $\sigma^2 = \|\Sigma\|_{\text{op}}$. More generally, any *strongly log-concave* distribution satisfies the Poincaré inequality:

Fact 3.2 ([BÉ85]). *Suppose that $p(x) = \exp(-\psi(x))$, where ψ is a function with strictly positive curvature: $\nabla^2\psi(x) \geq \frac{1}{\sigma^2}I$ for all $x \in \mathbb{R}^d$. Then p satisfies the Poincaré inequality with parameter σ .*

We recover the Gaussian case by letting $\psi(x) = \frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)$. Fact 3.2 was first established by Bakry and Émery [BÉ85] who actually prove a stronger *log-Sobolev* inequality for p .

Another important class is the family of distributions with bounded support. It cannot be the case that an arbitrary bounded distribution satisfies (3.1), because we have already seen that no discrete distribution can satisfy (3.1). However, it is always possible to add noise to the distribution that smooths out the support and allows (3.1) to hold. Specifically:

Fact 3.3 ([BGMZ18]). *Suppose that p is a distribution on \mathbb{R}^d whose support has radius at most R in ℓ_2 -norm. Let p_τ denote the result of adding Gaussian noise with variance $\tau^2 I$ to p . Then, if $\tau \geq 2R$, p_τ satisfies the Poincaré inequality with parameter $\tau\sqrt{e}$.*

So, we can always cause a bounded random variable to satisfy (3.1) by adding sufficiently large Gaussian noise. We remark that while this is very useful in low dimensions, in high dimensions the radius R typically grows as \sqrt{d} , in which case Fact 3.3 does not give very good bounds.

Composition rules. The Poincaré inequality is also preserved under products, sums, and uniformly continuous transformations. Specifically, we have the following:

Proposition 3.4. *The following composition rules hold:*

- If independent random variables X_1 and X_2 are σ_1 - and σ_2 -Poincaré, respectively, then the product distribution (X_1, X_2) is $\max(\sigma_1, \sigma_2)$ -Poincaré.
- If independent random variables X_1 and X_2 are σ_1 - and σ_2 -Poincaré, respectively, then $aX_1 + bX_2$ is $\sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}$ -Poincaré.
- If X is σ -Poincaré and $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ satisfies $\|\Phi(x') - \Phi(x)\|_2 \leq L\|x' - x\|_2$ for all $x, x' \in \mathbb{R}^d$, then $\Phi(X)$ is $(L\sigma)$ -Poincaré.

The above properties are all straightforward, but for completeness we prove them in Appendix B.

Implications of the Poincaré inequality. We end with some implications of (3.1), which in particular specify properties that any distribution satisfying (3.1) must have.

First, any distribution satisfying the Poincaré inequality has exponentially decaying tails. Specifically, the following is a well known fact:

Fact 3.5 (Tail Bound for Poincaré Distributions). *For any unit vector v , and X a σ -Poincaré random variable with mean μ , we have*

$$\mathbb{P}[|\langle v, x - \mu \rangle| \geq z] \leq 6 \exp(-z/\sigma). \quad (3.2)$$

More generally, any distribution satisfying the Poincaré inequality also satisfies a Lipschitz concentration property:

Fact 3.6 ([BL97]). *If p satisfies the Poincaré inequality with parameter σ and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies $\|\nabla f(x)\|_2 \leq L$ for all x , then $\mathbb{E}[\exp(f(x) - \mathbb{E}[f(x)])] \leq \frac{2+\sigma L}{2-\sigma L}$. In particular, $\mathbb{P}[|f(x) - \mathbb{E}[f(x)]| \geq z] \leq 6 \exp(-\frac{z}{\sigma L})$.*

Exponential concentration of Lipschitz functions (with weaker bounds than above) was first observed in [GM83]. Fact 3.6 generalizes the previous point on exponential concentration of linear functions, as can be seen by taking $f(x) = \langle v, x \rangle$.

Finally, and most important to our subsequent analysis, the gradient bound (3.1) implies an analogous bound for higher-order derivatives as well. Specifically, (3.1) can be re-written as saying that $\mathbb{E}[f(x)^2] \leq \sigma^2 \mathbb{E}[\|\nabla f(x)\|_2^2]$ whenever $\mathbb{E}[f(x)] = 0$. More generally, we have:

Fact 3.7 ([AW15]). *Suppose that p satisfies the Poincaré inequality with parameter σ . Then, if $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a function satisfying $\mathbb{E}_{x \sim p}[f(x)] = \mathbb{E}_{x \sim p}[\nabla f(x)] = \dots = \mathbb{E}_{x \sim p}[\nabla^{t-1} f(x)] = 0$, we have*

$$\mathbb{E}_{x \sim p}[f(x)^2] \leq C_t \sigma^{2t} \mathbb{E}_{x \sim p}[\|\nabla^t f(x)\|_F^2], \quad (3.3)$$

where C_t is a constant depending only on t .

For $t = 1$, we recover the usual Poincaré inequality. Thus Fact 3.7 can be interpreted as saying that if (3.3) holds for $t = 1$, it holds for all $t > 1$ as well.

Despite its simplicity, Fact 3.7 is actually a highly non-trivial consequence of the Poincaré inequality. It is a special case of results due to Adamczak and Wolff [AW15] (see Theorem 3.3 and the ensuing discussion therein); those results in turn build on work of Latała [Lat06].

In the next section, we will use Fact 3.7 to obtain a low-degree sum-of-squares proof of a sharp upper bound on the injective norms of moment tensors of arbitrary Poincaré distributions.

4 Certifying Injective Norms for Poincaré Distributions

Fact 3.5 shows an upper bound on $\mathbb{E}[\exp(\frac{1}{\sigma} \langle x, v \rangle)]$ for σ -Poincaré distributions, and hence in particular on the $2t$ th moments $\sup_{\|v\|_2 \leq 1} \mathbb{E}[\langle x, v \rangle^{2t}]$ for any t . The goal of this section is to show a low-degree sum-of-squares proof of this fact.

Specifically, the main goal is to show the following:

Theorem 4.1. *Let p be a zero-mean, σ -Poincaré distribution with $2t$ -th moment tensor $M_{2t}(p) = \mathbb{E}_{x \sim p}[(x - \mu)^{\otimes 2t}]$, where $\mu = \mathbb{E}_{x \sim p}[x]$ is the mean of p . Then, for all t , $M_{2t}(p)$ is $(2t, C_t \sigma)$ -SOS-certifiable:*

$$\langle M_{2t}(p), v^{\otimes 2t} \rangle \leq_{2t} C_t \sigma^{2t} \|v\|_2^{2t} \text{ for some universal constant } C_t. \quad (4.1)$$

Moreover, given $n \geq (2d \log(td/\delta))^t$ samples from p , with probability $1 - \delta$ the moment tensor \hat{M}_{2t} of the empirical distribution will also satisfy (4.1) (with a different constant C_t).

Recall that $p(v) \leq_{2t} q(v)$ means that there is a degree- $2t$ sum-of-squares proof that $q(v) - p(v) \geq 0$ (as a polynomial in v).

In the rest of this section, we will first warm up by proving Theorem 4.1 for $t = 1, 2$; the proof in these cases is standard for sum-of-squares experts, but will help to illustrate a few important ideas used in the sequel. Next, in Section 4.1 we will prove Theorem 4.1 for $t = 3$, which is no longer

standard and contains most of the ideas in the general case. Finally, we will prove the general case in Section 4.2. We leave the issue of finite-sample concentration to the very end.

For $t = 1$, Theorem 4.1 has a simple proof: $\langle M_2, v^{\otimes 2} \rangle \leq_2 \|M_2\|_{\text{op}} \|v\|_2^2 \leq_2 O(\sigma^2) \|v\|_2^2$ using Fact 2.2 and the fact that σ -Poincaré distributions have covariance $O(\sigma^2)$. This simply asserts that degree-2 sum-of-squares knows about spectral norm upper bounds.

For $t = 2$, a natural idea is to flatten M_4 (which is a $d \times d \times d \times d$ tensor) into a $d^2 \times d^2$ matrix $F(M_4)$, and obtain upper bounds in terms of the spectral norm of this flattened matrix.

Unfortunately, even for a Gaussian distribution this estimate can be off by a factor as large as the dimension d . Specifically, if $p = \mathcal{N}(0, \Sigma)$ is a Gaussian with variance Σ , then the flattening $F(M_4)$ is equal to $\text{vec}(\Sigma) \text{vec}(\Sigma)^\top + 2\Sigma \otimes \Sigma$, where $\text{vec}(\Sigma)$ flattens the d -dimensional matrix Σ to a d^2 -dimensional vector, and $(A \otimes B)_{i'j', i''j''} = A_{ij} B_{i''j''}$. (We skip the standard argument based on Isserlis' theorem.) This is problematic because it means that $\|F(M_4)\|_{\text{op}}^2 \geq \|\text{vec}(\Sigma)\|_2^2 = \|\Sigma\|_F^2$. If e.g. $\Sigma = I_d$ is the identity matrix, then $\|\Sigma\|_F^2 = d$, while we would hope to certify an upper bound of $O(1)$.

The key idea that allows us to get a (much) improved bound here is to observe that, as polynomials in v , $\langle v^{\otimes 2}, \text{vec}(\Sigma) \text{vec}(\Sigma)^\top v^{\otimes 2} \rangle = \langle v^{\otimes 2}, (\Sigma \otimes \Sigma) v^{\otimes 2} \rangle$. That is, the degree-4 polynomials defined by $\text{vec}(\Sigma) \text{vec}(\Sigma)^\top$ and $\Sigma \otimes \Sigma$ are equal—this allows us to “change the representation” for the same polynomial to go to a “representation” where the associated matrix has a smaller spectral norm. This fact has a simple sum-of-squares proof (it is sometimes referred to by saying that pseudo-distributions respect PPT symmetries) and allows us to now upper bound

$$\langle v^{\otimes 2}, F(M_4) v^{\otimes 2} \rangle = \langle v^{\otimes 2}, (\text{vec}(\Sigma) \text{vec}(\Sigma)^\top + 2\Sigma \otimes \Sigma) v^{\otimes 2} \rangle \quad (4.2)$$

$$= 3 \langle v^{\otimes 2}, (\Sigma \otimes \Sigma) v^{\otimes 2} \rangle \quad (4.3)$$

$$\leq_4 3 \|\Sigma \otimes \Sigma\|_{\text{op}}^2 \|v^{\otimes 2}\|_2^2 \quad (4.4)$$

$$= 3 \|\Sigma\|_{\text{op}}^4 \|v\|_2^4. \quad (4.5)$$

The above argument shows how one can exploit the symmetry properties of pseudo-distributions in order to certify strikingly better upper bounds on the maximum of the degree-4 polynomials associated with the moment tensors. This suggests a natural strategy for going beyond $t = 2$: write the moment tensor as a sum of (constantly many) terms and show that each term, as a polynomial, is equivalent to one where the canonical flattening as a matrix has a small spectral norm. This argument can (with much tedium) be made to work for small t s but can get unwieldy for large t s. However, the issue with the above argument is that it uses that the structure of the moment tensor was known to us. In our argument for arbitrary Poincaré distributions, we cannot rely on knowing the structure of the moment tensors and so will need a different proof technique.

Degree-2 proof for Poincaré distributions. To establish sum-of-squares bounds for general Poincaré distributions, we make use of Fact 3.7 from Section 3. Recall that Fact 3.7 states (for $t = 2$) that if $f(x)$ satisfies $\mathbb{E}[f(x)] = \mathbb{E}[\nabla f(x)] = 0$, then $\mathbb{E}[f(x)^2] \leq C \cdot \sigma^4 \cdot \mathbb{E}[\|\nabla^2 f(x)\|_F^2]$. We will define the polynomial $f_A(x) = \langle (x - \mu)(x - \mu)^\top - \Sigma, A \rangle$, where μ is the mean and $\Sigma = M_2(p)$ is the covariance of p .

Note that $\mathbb{E}[f_A(x)] = \mathbb{E}[\nabla f_A(x)] = 0$, while $\mathbb{E}[\|\nabla^2 f_A(x)\|_F^2] = \mathbb{E}[\|2A\|_F^2] = 4\|A\|_F^2$. Therefore, we have $\mathbb{E}[\langle (x - \mu)(x - \mu)^\top - \Sigma, A \rangle^2] \leq O(\sigma^4) \|A\|_F^2$ for all matrices A . This implies that for the tensor $F_2(x) = (x - \mu)(x - \mu)^\top - \Sigma$, we have $\langle \mathbb{E}[F_2(x)^{\otimes 2}], v^{\otimes 4} \rangle = \mathbb{E}[\langle F_2(x), v^{\otimes 2} \rangle^2] \leq_4 O(\sigma^4) \|v\|_2^4$ (by Fact 2.3).

Next note that $\mathbb{E}[F_2(x)^{\otimes 2}] = M_4 - \Sigma \otimes \Sigma$. Therefore, we have the sum-of-squares proof

$$\langle M_4, v^{\otimes 4} \rangle = \langle M_4 - \Sigma \otimes \Sigma, v^{\otimes 4} \rangle + \langle \Sigma \otimes \Sigma, v^{\otimes 4} \rangle \quad (4.6)$$

$$\leq_4 \mathcal{O}(\sigma^4) \|v\|_2^4 + \langle v^{\otimes 2}, (\Sigma \otimes \Sigma) v^{\otimes 2} \rangle \quad (4.7)$$

$$\leq_4 \mathcal{O}(\sigma^4) \|v\|_2^4 + \|\Sigma\|_{\text{op}}^2 \|v\|_2^4 \quad (4.8)$$

$$= \mathcal{O}(\sigma^4) \|v\|_2^4, \quad (4.9)$$

yielding the desired certificate for Poincaré distributions for $t = 2$.

The crucial step was the bound $\mathbb{E}[\langle F_2(x), A \rangle^2] = \mathbb{E}[\langle (x - \mu)(x - \mu)^\top - \Sigma, A \rangle^2] \leq \mathcal{O}(\sigma^4) \|A\|_F^2$, which bounds the covariance of $F_2(x)$ (when considered as a vector) and hence yields spectral information which is accessible to sum-of-squares via Fact 2.3. This inequality holds for σ -Poincaré distributions but *not* for arbitrary distributions with bounded 4th moments. The case of general $t > 2$ involves conjuring similar operator norm bounds for other polynomials in x which combine to yield bounds on the moment tensor M_{2t} . We will see this next for the case of $t = 3$.

4.1 The case of $t = 3$

We now handle the case of $t = 3$, assuming that we already know a bound for $t = 2$. (We will formally induct on t in the next section.) Specifically, in this section we will show:

Proposition 4.2. *Suppose that p is σ -Poincaré and we are given that $\langle M_2, v^{\otimes 2} \rangle \leq_2 \mathcal{O}(1) \cdot \sigma^2 \|v\|_2^2$ and $\langle M_4, v^{\otimes 4} \rangle \leq_4 \mathcal{O}(1) \cdot \sigma^4 \|v\|_2^4$. Then, $\langle M_6, v^{\otimes 6} \rangle \leq_6 \mathcal{O}(1) \cdot \sigma^6 \|v\|_2^6$.*

The key idea in the argument is to observe that Adamczak and Wolff's generalized Poincaré inequality (Fact 3.7) implies a sum of squares upper bound on the injective norm of a tensor related to M_6 .

Lemma 4.3. *Let p be a σ -Poincaré probability distribution. For any tensor A of order 3, let $f_A(x) = \langle (x - \mu)^{\otimes 3} - 3(x - \mu) \otimes M_2 - M_3, A \rangle$. Then, $\mathbb{E}_x[f_A^2(x)] \leq \mathcal{O}(1) \cdot \sigma^6 \|A\|_F^2$.*

Proof. The following is a direct computation:

$$\nabla f_A(x) = \langle 3(x - \mu)^{\otimes 2} \otimes I - 3M_2 \otimes I, A \rangle, \quad \text{and hence } \mathbb{E}[\nabla f_A(x)] = 0 \quad (4.10)$$

$$\nabla^2 f_A(x) = \langle 6(x - \mu) \otimes I \otimes I, A \rangle, \quad \text{and hence } \mathbb{E}[\nabla^2 f_A(x)] = 0 \quad (4.11)$$

$$\nabla^3 f_A(x) = \langle 6I \otimes I \otimes I, A \rangle = 6A, \quad \text{and hence } \mathbb{E}[\|\nabla^3 f_A(x)\|_F^2] = 36\|A\|_F^2. \quad (4.12)$$

Therefore, by Fact 3.7, we have $\mathbb{E}[f_A(x)^2] \leq \mathcal{O}(1) \cdot \sigma^6 \|A\|_F^2$ for all tensors A of order 3. \square

Define $F_3(x) = (x - \mu)^{\otimes 3} - 3(x - \mu) \otimes M_2 - M_3$. The conclusion of Lemma 4.3 together with Fact 2.3 implies that $\langle \mathbb{E}[F_3(x)^{\otimes 2}], v^{\otimes 6} \rangle = \mathbb{E}[\langle F_3(x)^{\otimes 2}, v^{\otimes 6} \rangle] \leq_6 \mathcal{O}(1) \cdot \sigma^6 \|v\|_2^6$.

To complete the proof, we will write $\langle M_6, v^{\otimes 6} \rangle$ (the polynomial we are trying to bound) as a linear combination of $\langle \mathbb{E}[F_3(x)^{\otimes 2}], v^{\otimes 6} \rangle$ together with other terms that we can bound by appealing to upper bounds on $\langle M_4, v^{\otimes 4} \rangle$ and $\langle M_2, v^{\otimes 2} \rangle$.

Proof of Proposition 4.2. First we compute $\mathbb{E}[F_3(x)^{\otimes 2}]$:

$$\mathbb{E}[F_3(x)^{\otimes 2}] = M_6 + 9M_2 \otimes M_2 \otimes M_2 - 6M_4 \otimes M_2 - M_3 \otimes M_3. \quad (4.13)$$

Since we know that $\langle \mathbb{E}[F_3(x)^{\otimes 2}], v^{\otimes 6} \rangle \leq_6 \mathcal{O}(1) \cdot \sigma^6 \|v\|_2^6$, (4.13) implies that

$$\langle M_6, v^{\otimes 6} \rangle = \langle \mathbb{E}[F_3(x)^{\otimes 2}], v^{\otimes 6} \rangle + \langle M_6 - \mathbb{E}[F_3^{\otimes 2}], v^{\otimes 6} \rangle \quad (4.14)$$

$$\leq_6 \mathcal{O}(1) \cdot \sigma^6 \|v\|_2^6 + 6\langle M_4, v^{\otimes 4} \rangle \langle M_2, v^{\otimes 2} \rangle + \langle M_3, v^{\otimes 3} \rangle^2 - 9\langle M_2, v^{\otimes 2} \rangle^3. \quad (4.15)$$

The final term $-9\langle M_2, v^{\otimes 2} \rangle^3$ is a negative sum of squares, so we can ignore it. On the other hand, we have

$$6\langle M_4, v^{\otimes 4} \rangle \langle M_2, v^{\otimes 2} \rangle + \langle M_3, v^{\otimes 3} \rangle^2 \stackrel{(i)}{\leq_6} 7\langle M_4, v^{\otimes 4} \rangle \langle M_2, v^{\otimes 2} \rangle \quad (4.16)$$

$$\stackrel{(ii)}{\leq_6} 7 \cdot (\mathcal{O}(1) \cdot \sigma^4 \|v\|_2^4) \cdot (\mathcal{O}(1) \cdot \sigma^2 \|v\|_2^2) = \mathcal{O}(1) \cdot \sigma^6 \|v\|_2^6. \quad (4.17)$$

Here we used two facts: (i) $\langle M_3, v^{\otimes 3} \rangle^2 \leq_6 \langle M_2, v^{\otimes 2} \rangle \langle M_4, v^{\otimes 4} \rangle$ (by Cauchy-Schwarz) and (ii) Fact 2.4 on multiplying together sum-of-squares inequalities.

As we have bounded all terms by $\mathcal{O}(1) \cdot \sigma^6 \|v\|_2^6$, the proof is complete. \square

4.2 The case of arbitrary t

We now generalize the argument from the previous section to arbitrary t s.

As in the case of $t = 3$ above, we will first come up with a polynomial of degree $2t$ that we can upper bound directly using the generalized Poincaré inequality. For this purpose, we define the order- t tensor $F_t(x)$ generalizing F_3 from the previous section.

Let \mathcal{T}_t denote the set of all tuples (i_0, i_1, \dots, i_r) of integers such that $i_0 \geq 0$ and $i_s \geq 2$ for $s > 0$, and $i_0 + \dots + i_r = t$. We will take

$$F_t(x) \stackrel{\text{def}}{=} \sum_{(i_0, \dots, i_r) \in \mathcal{T}_k} (-1)^r \binom{k}{i_0 \dots i_r} (x - \mu)^{\otimes i_0} \otimes M_{i_1} \otimes \dots \otimes M_{i_r}. \quad (4.18)$$

While (4.18) may seem mysterious, $F_t(x)$ is in fact the unique tensor such that $\nabla^t F_t(x) = I$ and $\mathbb{E}[\nabla^s F_t(x)] = 0$ for $s < t$.

We verify the latter property in the following Lemma 4.4 (we omit the proof of uniqueness). For any order- t tensor A , F_t defines a polynomial $f_A(x) = \langle F_t(x), A \rangle$. We will show that the partial derivatives of f_A (w.r.t x) of all orders $0, 1, \dots, t-1$ have expectation zero, so that we can apply the higher-order Poincaré inequality from Fact 3.7.

Lemma 4.4. *For every order- t A and for every $j = 0, \dots, t-1$, $\mathbb{E}[\langle \nabla^j F_t(x), A \rangle] = 0$.*

Proof. We show the following structural property of the partial derivative tensors of F for every x :

Claim 4.5. $\langle \nabla F_t(x), A \rangle = t \langle F_{t-1}(x) \otimes I, A \rangle$ for every symmetric tensor A .

Proof of Claim. We have

$$\nabla \langle F_t(x), A \rangle = \sum_{(i_0, \dots, i_r) \in \mathcal{T}_t} (-1)^r \binom{t}{i_0 \dots i_r} \langle \nabla (x - \mu)^{\otimes i_0} \otimes M_{i_1} \otimes \dots \otimes M_{i_r}, A \rangle \quad (4.19)$$

$$= \sum_{(i_0, \dots, i_r) \in \mathcal{T}_t} (-1)^r i_0 \binom{k}{i_0 \dots i_r} \langle (x - \mu)^{\otimes (i_0-1)} \otimes I \otimes M_{i_1} \otimes \dots \otimes M_{i_r}, A \rangle \quad (4.20)$$

$$\stackrel{(i)}{=} t \sum_{(i_0-1, \dots, i_r) \in \mathcal{T}_{t-1}} (-1)^r \binom{k}{i_0-1 \dots i_r} \langle (x - \mu)^{\otimes (i_0-1)} \otimes I \otimes M_{i_1} \otimes \dots \otimes M_{i_r}, A \rangle \quad (4.21)$$

$$\stackrel{(ii)}{=} t \langle F_{t-1}(x) \otimes I, A \rangle, \quad (4.22)$$

In (i), we fold the factor of i_0 into the multinomial coefficient, and in (ii) we use the symmetries of A to permute tensor modes. \square

By Claim 4.5, it suffices to show that $\mathbb{E}_x[\langle F_t(x), A \rangle] = 0$ for $t > 0$ (since the derivatives of F_t are simply of the form F_s for $s < t$). Note that

$$\mathbb{E}_x[\langle F_t(x), A \rangle] = \sum_{(i_0, \dots, i_r) \in \mathcal{T}_k} (-1)^r \binom{k}{i_0 \dots i_r} \langle M_{i_0} \otimes \dots \otimes M_{i_r}, A \rangle. \quad (4.23)$$

Now we create an involution within the terms of \mathcal{T}_k that matches terms of equal magnitudes but opposite signs. This will establish that the RHS of (4.23) vanishes.

For every term (i_0, \dots, i_r) with $i_0 \neq 0$, match it with the term $(0, i_0, \dots, i_r)$. Conversely, if $i_0 = 0$, match it with (i_1, \dots, i_r) . This is an involution which preserves $\binom{k}{i_0 \dots i_r}$ and $M_{i_0} \otimes \dots \otimes M_{i_r}$ but negates $(-1)^r$, so all terms will cancel. The only exception is if $i_0 = 1$ (then $(0, i_0, \dots, i_r) \notin \mathcal{T}_k$ so the involution fails), but this term is already zero because $M_1 = \mathbb{E}[x - \mu] = 0$. \square

With Lemma 4.4 in hand, we now know that the partial derivatives of $f_A(x) = \langle F_t(x), A \rangle$ are all mean-zero, and so we can apply the higher-order Poincaré inequality (Fact 3.7).

We therefore have $\mathbb{E}[f_A(x)^2] \leq (t!)^2 C_t \sigma^{2t} \|A\|_F^2$ for the constant C_t for which Fact 3.7 holds. Using Fact 2.3 as before, we obtain the following corollary asserting that $\mathbb{E}[F_t(x)^{\otimes 2}]$ has a sum-of-squares upper bound:

Corollary 4.6. *For all t , we have*

$$\langle \mathbb{E}[F_t(x)^{\otimes 2}], v^{\otimes 2t} \rangle = \mathbb{E}[\langle F_t(x), v^{\otimes t} \rangle^2] \leq_{2t} (t!)^2 C_t \sigma^{2t} \|v\|_2^{2t}. \quad (4.24)$$

Proving Theorem 1.1. We are now ready to prove Theorem 1.1. We will first show that Theorem 1.1 holds in the infinite-data limit, and attend to finite-sample concentration at the end.

As in the case of $t = 3$, the strategy is to write $\langle M_{2t}, v^{\otimes 2t} \rangle$ as a combination of $\langle \mathbb{E}[F_t(x)^{\otimes 2}], v^{\otimes 2t} \rangle$ together with terms that we can upper bound by recursively relying on estimates from smaller values of t . To aid in this, we use the following generalization of Hölder's inequality:

Lemma 4.7. *For every collection of non-negative integers i_0, \dots, i_r that sum to $2t$ and are each at most $2t - 2$, $\langle M_{i_0} \otimes \dots \otimes M_{i_r}, v^{\otimes 2t} \rangle \leq_{2t} \langle M_2 \otimes M_{2t-2}, v^{\otimes 2t} \rangle$.*

We will simply use Lemma 4.7 to replace all lower-order terms with terms of the form $\langle M_2 \otimes M_{2t-2}, v^{\otimes 2t} \rangle$. The proof of Lemma 4.7 involves providing sum-of-squares proofs for a number of standard polynomial inequalities, and is deferred to Appendix A.

We will next induct on t . The base case $t = 1$ was already given above. For $t > 1$, we have

$$\langle M_{2t}, v^{\otimes 2t} \rangle = \mathbb{E}[\langle (x - \mu)^{\otimes t}, v^{\otimes t} \rangle^2] \quad (4.25)$$

$$\stackrel{(i)}{=} \mathbb{E}[\langle F_t(x), v^{\otimes t} \rangle^2] - 2\mathbb{E}[\langle (x - \mu)^{\otimes t}, v^{\otimes t} \rangle \langle F_t(x) - (x - \mu)^{\otimes t}, v^{\otimes t} \rangle] - \mathbb{E}[\langle F_t(x) - (x - \mu)^{\otimes t}, v^{\otimes t} \rangle^2] \quad (4.26)$$

$$\stackrel{(ii)}{\leq_{2t}} (t!)^2 C_t \sigma^{2t} \|v\|_2^{2t} - 2\mathbb{E}[\langle (x - \mu)^{\otimes t}, v^{\otimes t} \rangle \langle F_t(x) - (x - \mu)^{\otimes t}, v^{\otimes t} \rangle]. \quad (4.27)$$

Here (i) is direct algebra while (ii) is applying Corollary 4.6, as well as the fact that $-\mathbb{E}[\langle F_t(x) - (x - \mu)^{\otimes t}, v^{\otimes t} \rangle^2] \leq_{2t} 0$.

We next want to bound the $-2\mathbb{E}[\langle (x - \mu)^{\otimes t}, v^{\otimes t} \rangle \langle F_t(x) - (x - \mu)^{\otimes t}, v^{\otimes t} \rangle]$ term. Note that $F_t(x) - (x - \mu)^{\otimes t}$ is simply $F_t(x)$ without its leading term. Therefore, recalling the definition (4.18)

of $F_t(x)$, we will let $\mathcal{T}'_t = \mathcal{T}_t \setminus \{(t)\}$ (since the tuple (t) is the one generating the $(x - \mu)^{\otimes t}$ term in F_t). Then we have

$$-2\mathbb{E}[\langle (x - \mu)^{\otimes t}, v^{\otimes t} \rangle \langle (F_t(x) - (x - \mu)^{\otimes t}), v^{\otimes t} \rangle] \quad (4.28)$$

$$= -2 \left\langle \sum_{(i_0, \dots, i_r) \in \mathcal{T}'_t} (-1)^r \binom{t}{i_0 \dots i_r} M_{i_0+t} \otimes M_{i_1} \otimes \dots \otimes M_{i_r}, v^{\otimes 2t} \right\rangle \quad (4.29)$$

$$\stackrel{(i)}{\leq}_{2t} 2 \left\langle \sum_{(i_0, \dots, i_r) \in \mathcal{T}'_t} \binom{t}{i_0 \dots i_r} M_{2t-2} \otimes M_2, v^{\otimes 2t} \right\rangle \quad (4.30)$$

$$= 2 \left\langle \sum_{(i_0, \dots, i_r) \in \mathcal{T}'_t} \binom{t}{i_0 \dots i_r} \right\rangle \langle M_{2t-2} \otimes M_2, v^{\otimes 2t} \rangle. \quad (4.31)$$

Here (i) is by Lemma 4.7. We will bound the sum with a combinatorial argument. We can interpret the sum as the number of ways of splitting $\{1, \dots, t\}$ into some number r of sets such that all sets but the first have size at least 2. Since $r \leq t/2 + 1$, this is bounded above by the number of ways of splitting $\{1, \dots, t\}$ into $t/2 + 1$ (possibly empty) sets. But this is just $(t/2 + 1)^t$, as each element can freely go into one of the $t/2 + 1$ sets. Therefore, (4.31) is at most $2(t/2 + 1)^t \langle M_{2t-2} \otimes M_2, v^{\otimes 2t} \rangle$. Plugging back into (4.27), we get

$$\langle M_{2t}, v^{\otimes 2t} \rangle \leq_{2t} (t!)^2 C_t \sigma^{2t} \|v\|_2^{2t} + 2(t/2 + 1)^t \langle M_{2t-2} \otimes M_2, v^{\otimes 2t} \rangle. \quad (4.32)$$

Now by the inductive hypothesis, we have both $\langle M_2, v^{\otimes 2} \rangle \leq_2 C'_2 \sigma^2 \|v\|_2^2$ and $\langle M_{2t-2}, v^{\otimes (2t-2)} \rangle \leq_{2t-2} C'_{2t-2} \sigma^{2t-2} \|v\|_2^{2t-2}$ for some constants C'_2 and C'_{2t-2} . Therefore, $\langle M_{2t-2} \otimes M_2, v^{\otimes 2t} \rangle = \langle M_{2t-2}, v^{\otimes (2t-2)} \rangle \langle M_2, v^{\otimes 2} \rangle \leq_{2t} C'_2 C'_{2t-2} \sigma^{2t} \|v\|_2^{2t}$ (by Fact 2.4). We therefore see that $\langle M_{2t}, v^{\otimes 2t} \rangle \leq_{2t} C'_2 \sigma^{2t} \|v\|_2^{2t}$, where we can take $C'_2 = (t!)^2 C_t + 2(t/2 + 1)^t C'_2 C'_{2t-2}$. This completes the induction.

Finite-sample concentration. To finish the proof of Theorem 4.1, it remains to establish finite-sample concentration. The key observation is that, as long as Corollary 4.6 holds for $s = 1, \dots, t$, then all of the above steps go through. But Corollary 4.6 relies only on a bound on the maximum eigenvalue of the covariance of $F_t(x)$ (when $F_t(x)$ is flattened to a vector). It therefore suffices to bound this maximum eigenvalue in finite samples.

It is tempting to apply standard matrix concentration bounds (such as the matrix Chernoff inequality; see Theorem 5.5.1 of [Tro15]); alas, $F_t(x)$ is too heavy-tailed for this to be valid (i.e., it is unbounded and does not even have exponential moments as required for most common matrix concentration bounds). We must instead appeal to the Matrix Rosenthal inequality (Corollary 7.4 of [MJC⁺14]). In our context, if we let $\text{vec}(F_t(x))$ denote the flattening of $F_t(x)$ to a vector and $\Sigma = \mathbb{E}[\text{vec}(F_t(x)) \text{vec}(F_t(x))^\top]$ denote the covariance of this vector (and $\hat{\Sigma}$ denote the empirical covariance given n samples), then the Matrix Rosenthal inequality states that for any $p \geq 1.5$, we have

$$\mathbb{E}[\text{tr}(\hat{\Sigma}^{2p})]^{1/2p} \leq \left(\sqrt{\text{tr}(\Sigma^{2p})}^{1/2p} + \sqrt{\frac{4p-2}{n}} n^{1/4p} \mathbb{E}[\|F_t(x)\|_F^{4p}]^{1/4p} \right)^2 \quad (4.33)$$

$$\leq 2(\text{tr}(\Sigma^{2p})^{1/2p} + \frac{4p-2}{n^{1-1/2p}} \mathbb{E}[\|F_t(x)\|_F^{4p}]^{1/2p}). \quad (4.34)$$

Using a stronger version of Adamczak and Wolff's result (given as Theorem 3.3 of [AW15]) we obtain the bound $\mathbb{E}[\|F_t(x)\|_F^{4p}] \leq (C_t d^{t/2} p^t \sigma^t)^{4p}$ for some constant C_t . We also have $\text{tr}(\Sigma^{2p})^{1/2p} \leq$

$d^{t/2p} \|\Sigma\|_{\text{op}} \leq C_t d^{t/2p} \sigma^{2t}$ as $\|\Sigma\|_{\text{op}}$ is bounded by Corollary 4.6 (note that here and below the constant C_t changes in each instance). Plugging into the inequality above, we obtain

$$\mathbb{E}[\text{tr}(\hat{\Sigma}^{2p})]^{1/2p} \leq C_t \sigma^{2t} (d^{t/2p} + d^t p^{2t+1} n^{1/2p-1}). \quad (4.35)$$

We will take $p^* = t \log(d/\delta) + \frac{1}{2}$ which yields

$$\mathbb{E}[\text{tr}(\hat{\Sigma}^{2p^*})]^{1/2p^*} \leq C_t \sigma^{2t} (1 + d^t \log^t(d/\delta) n^{1/2p^*-1}). \quad (4.36)$$

This is bounded so long as $n \geq ((d \log(d/\delta))^{\frac{2p^*}{2p^*-1}})^t$. For the value of p^* above one can check that $(d \log(d/\delta))^{\frac{2p^*}{2p^*-1}} \leq 2d \log(d/\delta)$, so the above is bounded by $C_t \sigma^{2t}$ so long as $n \geq (2d \log(d/\delta))^t$.

We thus have a bound on $\mathbb{E}[\text{tr}(\hat{\Sigma}^{2p^*})]$ which we would like to turn into a high-probability bound on $\|\hat{\Sigma}\|_{\text{op}}$. For this we make use of the matrix Chebyshev bound (Proposition 6.2 of [MJC⁺14]), which in our case implies that $\mathbb{P}[\|\hat{\Sigma}\|_{\text{op}} \geq 2\mathbb{E}[\text{tr}(\hat{\Sigma}^{2p^*})]^{1/2p^*}] \leq 2^{-p^*} \leq \delta$. Union bounding over the t instances where we must invoke Corollary 4.6 completes the proof.

5 Robust Clustering

In this section, we give our algorithms for robust clustering under separation assumptions and for robust mean estimation. The first main result of this section is that one can efficiently recover a good clustering of data, even in the presence of outliers, for arbitrary data that is $(2t, B)$ -SOS-certifiable (i.e., where the $2t$ th moments are bounded for all pseudodistributions on the sphere).

Theorem 5.1. *Suppose $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ satisfies the following clusterability condition: It can be partitioned into sets I_1, \dots, I_k , where I_j has size $\alpha_j n$, together with an ε fraction of arbitrary outliers, where $\varepsilon = 1 - (\alpha_1 + \dots + \alpha_k)$. Furthermore, $\varepsilon \leq \frac{\alpha}{8}$, where $\alpha = \min_j \alpha_j$, and the sets I_j satisfy:*

1. For every j , I_j is SOS-certifiable around its mean $\mu_j = \mathbb{E}_{x \sim I_j}[x]$:

$$\left\| \mathbb{E}_{x \sim I_j} [(x - \mu_j)^{\otimes 2t}] \right\|_{\text{SOS}_{2t}} \leq B. \quad (5.1)$$

2. For every i, j , the means of distinct clusters are well-separated:

$$\|\mu_i - \mu_j\| \geq C_{\text{sep}} B \alpha^{-1/t}, \quad (5.2)$$

with $C_{\text{sep}} \geq C_0$ (with C_0 an absolute constant).

Then, there is a $(nd)^{O(t)}$ time algorithm that outputs $\hat{\mu}_1, \dots, \hat{\mu}_k$ satisfying for each $j = 1, \dots, k$,

$$\|\hat{\mu}_j - \mu_j\|_2 \leq B \cdot \mathcal{O} \left(\frac{\varepsilon}{\alpha} + C_{\text{sep}}^{-2t} \right)^{\frac{2t-1}{2t}}. \quad (5.3)$$

Remark 5.2. Note that if $\varepsilon \geq \alpha$, then the number of outliers can be as large as the smallest cluster. In that case, it is information theoretically impossible, in general, to find a unique list of k correct means. Whenever $\varepsilon \leq \alpha/8$, Theorem 5.1 gives a non-trivial guarantee on the recovered means.

Remark 5.3. As explained in Section 1, the dependence of the error on the separation C_{sep} is also information-theoretically necessary, even in one dimension. In particular, consider two clusters drawn from distributions with slightly overlapping tails; then it is impossible to tell whether a point in the overlap should come from one cluster or the other, which will lead to small but non-zero errors in the estimated means.

The certified bounded moment condition (5.1) is satisfied by data generated from a large class of general mixture models. As a corollary of Theorem 5.1, we obtain results for learning means in general mixture models (see Corollary 1.3).

Our second result is outlier-robust mean estimation where an ε fraction of the input points are arbitrary outliers.

Theorem 5.4. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be such that there exists an unknown subset $I \subseteq [n]$ of size $(1 - \varepsilon)n$ satisfying $\|\mathbb{E}_{x \sim I}[(x_i - \mu)^{\otimes 2t}]\|_{\text{SOS}_{2t}} \leq B$, where $\mu = \mathbb{E}_{x \sim I}[x]$. If $\varepsilon \leq \frac{1}{4}$, then there is an algorithm that runs in time $(nd)^{O(t)}$ and outputs an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 \leq O(B\varepsilon^{1-1/2t})$.*

Theorem 5.4 is in fact a corollary of Theorem 5.1 (in the special case of a single cluster), but we state it separately for emphasis (and because Theorem 5.1 actually requires Theorem 5.4 in its proof). The error $\varepsilon^{1-1/2t}$ interpolates between existing results which achieve error $\sqrt{\varepsilon}$ for distributions with bounded covariance, and those achieving $\tilde{O}(\varepsilon)$ for e.g. Gaussian distributions.

Finally, we can obtain results in an even more extreme setting, where all but an α fraction of the input points are outliers, for α potentially smaller than $\frac{1}{2}$. This corresponds to the robust learning setting proposed in [CSV17], and our algorithm works in the *list-decodable learning* model [BBV08] in which a short list of $O(1/\alpha)$ candidate means is allowed to be output. (Note that this is information-theoretically necessary if $\alpha < \frac{1}{2}$.)

Theorem 5.5. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be such that there exists a subset $I \subseteq [n]$ of size αn that satisfies $\|\mathbb{E}_{x \sim I}(x_i - \mu)^{\otimes 2t}\|_{\text{SOS}_{2t}} \leq B$ for some $\mu \in \mathbb{R}^d$. Then, there is an algorithm that runs in time $(nd)^{O(t)}$ and outputs estimates $\hat{\mu}_1, \dots, \hat{\mu}_h$ with $h \leq \frac{4}{\alpha}$ and $\min_{j=1}^h \|\hat{\mu}_j - \mu\|_2 \leq O(B/\alpha^{1/t})$.*

Theorem 5.5 implies results for robust clustering, as we can think of each component I_j of the cluster as being the set I , and Theorem 5.5 then says we will output a list of $4/\alpha$ candidate means such that the mean of every cluster I_j is within $O(B/\alpha^{1/t})$ of a candidate mean $\hat{\mu}_j$. This is weaker than the guarantee of Theorem 5.1, but holds even when the clusters are not well-separated.

In fact, while Theorem 5.1 may appear much stronger in the well-separated setting, it will follow as a basic extension of the ideas in Theorems 5.5 and 5.4. Most of the rest of this section will be devoted to proving these two theorems, with Theorem 5.1 handled at the end.

5.1 Basic Clustering Relaxation

Key to our algorithmic results is a natural hierarchy of convex relaxations for recovering an underlying clustering. In what follows, we will refer to this as the basic clustering relaxation.

The idea is as follows. Assume (as in our results above) that there is some set I such that $\|\mathbb{E}_{x \sim I}[(x - \mu)^{\otimes 2t}]\|_{\text{SOS}_{2t}} \leq B$. This is the same as saying that for all pseudodistributions $\xi(v)$ over the unit sphere, $\frac{1}{|I|} \sum_{x \in I} \tilde{\mathbb{E}}_{\xi(v)}[\langle x - \mu, v \rangle^{2t}] \leq B^{2t}$. Motivated by this, we might seek w_1, \dots, w_n such that

$$\frac{1}{n} \sum_{i=1}^n \tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - w_i, v \rangle^{2t}] \text{ is small for all pseudodistributions } \xi \text{ over the unit sphere.} \quad (5.4)$$

The reason we allow distinct w_i is because in general the data might consist of multiple clusters and so we want to allow flexibility for the w_i to fit more than a single cluster at once.

However, as stated, the trivial solution $w_i = x_i$ will always have zero cost. Intuitively, the reason is that the w_i are free to completely overfit the x_i . We would like to impose an additional

penalty term to keep w_i from overfitting x_i too much. In fact, the right metric of overfitting turns out to be

$$\frac{1}{|I|} \sum_{i \in I} \langle w_i - \mu, w_i \rangle^{2t}, \quad (5.5)$$

as we will see below. We cannot directly control $w_i - \mu$ (as we do not know μ), but we can observe that $w_i - \mu = (w_i - x_i) + (x_i - \mu)$, and that both $w_i - x_i$ and $x_i - \mu$ have small moments with respect to any pseudodistribution ξ (by the constraint (5.4), and by our assumption on μ). So, we could conservatively try to control $\langle z_i, w_i \rangle^{2t}$ for *all possible* sets of points z_i that have small moments (which would in particular control $w_i - x_i$ and $x_i - \mu$).

This motivates us to add the following additional constraint: for any z_1, \dots, z_n such that $\sum_{i=1}^n \tilde{\mathbb{E}}_{\xi'(v)}[\langle z_i, v \rangle^{2t}] \leq 1$ for all pseudodistributions ξ' on the sphere, we ask that $\sum_{i=1}^n \langle z_i, w_i \rangle^{2t}$ be small (again, we think of the z_i as standing in for the points $w_i - x_i$ and $x_i - \mu$). This constraint on the w_i is not convex, but we can take a sum-of-squares relaxation by replacing the z_i with pseudodistributions $\zeta_i(z_i)$. This turns out to ask that

$$\sum_{i=1}^n \tilde{\mathbb{E}}_{\zeta_i(z_i)}[\langle z_i, w_i \rangle^{2t}] \text{ is small whenever } \sum_{i=1}^n \tilde{\mathbb{E}}_{\zeta_i(z_i)}[z_i^{\otimes 2t}] \preceq_{\text{sos}} \mathcal{I}, \quad (5.6)$$

where \mathcal{I} is the order- $2t$ identity tensor and $T_1 \preceq_{\text{sos}} T_2$ means that $\langle T_1, v^{\otimes 2t} \rangle \leq_{\text{sos}} \langle T_2, v^{\otimes 2t} \rangle$ (as a polynomial in v).

The basic clustering relaxation, defined in Algorithm 1 below, asks to either find w_1, \dots, w_n such that both (5.4) and (5.6) are small, or else to find dual certificates ξ, ζ proving that they cannot be small.

Algorithm 1 Basic Clustering Relaxation

- 1: Input: $X = x_1, \dots, x_n \in \mathbb{R}^d$, weights $c_1, \dots, c_n \in [0, 1]$, multiplier λ , threshold Γ .
- 2: Find either w_1, \dots, w_n such that

$$\sum_{i=1}^n c_i (\tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - w_i, v \rangle^{2t}] + \lambda \tilde{\mathbb{E}}_{\zeta_i(z_i)}[\langle w_i, z_i \rangle^{2t}]) \leq 2\Gamma \quad (5.7)$$

for all pseudodistributions ξ over the unit sphere and ζ_1, \dots, ζ_n with $\sum_{i=1}^n \tilde{\mathbb{E}}_{\zeta_i(z_i)}[z_i^{\otimes 2t}] \preceq_{\text{sos}} \mathcal{I}$, or else find $\xi, \zeta_{1:n}$ such that the expression in (5.7) is at least Γ for all $w_{1:n}$.

Note that Algorithm 1 is basically asking to implement an approximate separation oracle for the expression in (5.7). The weights c_1, \dots, c_n will be used later to downweight outliers in the data.

It is easy to show that Algorithm 1 specifies a convex primal-dual problem and can be solved in polynomial time. We give an argument in Section C of the Appendix for completeness:

Lemma 5.6 (Solving Basic Clustering Relaxation). *There is a polynomial time algorithm that solves the Basic Clustering Relaxation in time $(nd)^{O(t)}$.*

The full algorithm uses the relaxation (5.7) but must handle two additional issues. The first is *outliers*, which can prevent (5.7) from being small, and must be removed in a separate step. The second is the need for *re-clustering*. This second issue arises because (5.7) is not translation-invariant, and in particular the recovery error after running (5.7) will depend on the ℓ_2 -norm $r = \|\mu\|_2$ of μ . We can obtain improved bounds by clustering the w_i output by (5.7), and then re-running the algorithm on each cluster with a smaller value of r . This is similar to the re-clustering idea

in [CSV17], but we obtain a much simpler proof by making use of the recent idea of *resilience* introduced in [SCV18].

We analyze and give pseudocode for the outlier removal and re-clustering steps in Sections 5.2 and 5.4 below. The output of the outlier removal algorithm already satisfies a basic but coarse error bound, given as Proposition 5.7.

5.2 Outlier Removal: Basic Bound

The outlier removal algorithm is given in Algorithm 2. We maintain weights c_i on the points x_i , and run Algorithm 1 to attempt to find points $w_{1:n}^*$. If we fail, then we downweight c_i according to the value of $\tau_i^* = \min_w \tau_i(w)$, where $\tau_i(w) = \tilde{\mathbb{E}}[\langle x_i - w, v \rangle^{2t}] + \lambda \tilde{\mathbb{E}}[\langle w, z_i \rangle^{2t}]$ (line 10). This is intuitive because, whenever Algorithm 1 fails to output $w_{1:n}$, it instead outputs a dual certificate ξ, ζ such that $\sum_{i=1}^n c_i \tau_i^*$ is large. Thus, intuitively, points i with a large value of τ_i^* are responsible for (5.7) being large, and should be downweighted.

Algorithm 2 Outlier Removal Algorithm

- 1: Input: $x_1, \dots, x_n, B, \alpha$, and upper bound r on $\|\mu\|_2$.
 - 2: Initialize $c \leftarrow [1; \dots; 1] \in \mathbb{R}^n$
 - 3: Set $\lambda \leftarrow \alpha n (B/r)^{2t}$
 - 4: **while true do**
 - 5: Run Algorithm 1 with threshold $\Gamma = 4(nB^{2t} + \lambda r^{2t}/\alpha)$ to obtain either $w_{1:n}^*$ or $\xi^*, \zeta_{1:n}^*$.
 - 6: **if** $w_{1:n}^*$ **are obtained then**
 - 7: **return** $w_{1:n}^*, c_{1:n}$
 - 8: **else**
 - 9: Let $\tau_i^* \leftarrow \min_w \tau_i(w)$, where $\tau_i(w) = \tilde{\mathbb{E}}_{\xi^*(v)}[\langle x_i - w, v \rangle^{2t}] + \lambda \tilde{\mathbb{E}}_{\zeta_i^*(z_i)}[\langle w, z_i \rangle^{2t}]$.
 - 10: $c_i \leftarrow c_i(1 - \tau_i^*/\tau_{\max})$ for all i , where $\tau_{\max} = \max_{i=1}^n \tau_i^*$.
 - 11: **end if**
 - 12: **end while**
-

We analyze Algorithm 2 in two steps. First, we show that if the value of (5.7) is large (and hence $\xi^*, \zeta_{1:n}^*$ are obtained), then the re-weighting step (line 10) downweights bad points much more than it downweights good points. Second, we show that if the value of (5.7) is small, then the returned w_i^* constitute a good clustering (such that one of the clusters is centered close to the true mean μ).

Formally, we will show:

Proposition 5.7. *Suppose that there is a set $I \subseteq [n]$ of size αn such that $\|\frac{1}{|I|} \sum_{i \in I} (x_i - \mu)^{\otimes 2t}\|_{\text{sos}_{2t}} \leq B$, and $\|\mu\|_2 \leq r$. Then the output $w_{1:n}, c_{1:n}$ of Algorithm 2 satisfies the following property:*

$$\frac{1}{|I|} \sum_{i \in I} c_i \|w_i^* - \mu\|_2^{4t} \leq \mathcal{O}((4Br)^{2t}/\alpha^2) \text{ and } \frac{1}{|I|} \sum_{i \in I} (1 - c_i) \leq \frac{1 - \alpha}{3}. \quad (5.8)$$

In particular, at least a $\frac{1}{2}$ fraction of the $i \in I$ satisfy $c_i \geq \frac{1}{4}$ and $\|w_i^ - \mu\|_2^2 \leq \mathcal{O}(rB\alpha^{-1/t})$. Moreover, if $\alpha = 1 - \varepsilon$, with $\varepsilon \leq \frac{1}{2}$, then at least $1 - \varepsilon$ of the $i \in I$ satisfy $c_i \geq \frac{1}{4}$ and $\|w_i^* - \mu\|_2^2 \leq \mathcal{O}(rB\varepsilon^{-1/2t})$.*

Proof. Given (5.8), the remainder of Proposition 5.7 follows from repeated application of Markov's inequality. So, we will focus on establishing (5.8), splitting into cases based on the value of (5.7).

In particular, for the chosen threshold $\Gamma = 4(nB^{2t} + \lambda r^{2t}/\alpha)$, note that either Algorithm 1 outputs $\xi^*, \zeta_{1:n}^*$ showing that (5.7) is large (at least Γ), or it outputs $w_{1:n}^*$ such that (5.7) is small (at most 2Γ).

Case 1: (5.7) is large. Suppose that (5.7) is large, and hence in particular $\sum_{i=1}^n c_i \tau_i^* \geq 4(nB^{2t} + \lambda r^{2t}/\alpha)$. We start by showing that $\sum_{i \in I} \tau_i^*$ is much smaller than this. Indeed, we have $\sum_{i \in I} \tau_i^* \leq \sum_{i \in I} \tau_i(\mu)$. To bound $\tau_i(\mu)$, first note that $\sum_{i \in I} \tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - \mu, v \rangle^{2t}] \leq |I|B^{2t}$, since we are assuming a sum-of-squares certificate for I . Also, $\sum_{i \in I} \tilde{\mathbb{E}}_{\zeta_i(z_i)}[\langle \mu, z_i \rangle^{2t}] \leq \langle \mu^{\otimes 2t}, \mathcal{I} \rangle \leq r^{2t}$. Therefore,

$$\sum_{i \in I} c_i \tau_i^* \leq \sum_{i \in I} \tau_i(\mu) \leq |I|B^{2t} + \lambda r^{2t}. \quad (5.9)$$

Now, since $\sum_{i=1}^n c_i \tau_i^* \geq 4(nB^{2t} + \lambda r^{2t}/\alpha)$ by assumption, the average of τ_i^* over $\{1, \dots, n\}$ is more than 4 times larger than the average of τ_i^* over I (since $|I| \geq \alpha n$). Let c_i and c'_i be the values of the weights before and after the update on line 10. We have, for any set S ,

$$\sum_{i \in S} c_i - c'_i = \frac{1}{\tau_{\max}} \sum_{i \in S} c_i \tau_i^*. \quad (5.10)$$

Therefore, the amount that the weights in a set S decrease is proportional to $\sum_{i \in S} c_i \tau_i^*$. Since the average over I is at most $\frac{1}{4}$ the average over $\{1, \dots, n\}$, this means that the weights in I decrease at most $\frac{1}{4}$ as fast as the weights overall. In particular, at the end of the algorithm we have $\sum_{i \in I} (1 - c_i) \leq \frac{|I|}{4n} (\sum_{i \in I} (1 - c_i) + \sum_{i \notin I} (1 - c_i))$. Re-arranging yields $\frac{1}{|I|} \sum_{i \in I} (1 - c_i) \leq \frac{1 - |I|/n}{4 - |I|/n} \leq \frac{1 - |I|/n}{3}$. In particular, at most $\frac{1 - \alpha}{3}$ of the weight of the good c_i is removed at any stage in the algorithm (and hence in particular we eventually end up in case 2).

Case 2: (5.7) is small. Note that λ is chosen so that $8(nB^{2t} + \lambda r^{2t}/\alpha) = 16nB^{2t}$. Therefore, if (5.7) is small, then $\sum_{i \in I} c_i \tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - w_i^*, v \rangle^{2t}] \leq 16nB^{2t}$ for all pseudodistributions $\xi(v)$ over the unit sphere, and by assumption $\sum_{i \in I} c_i \tilde{\mathbb{E}}_{\xi(v)}[\langle x_i - \mu, v \rangle^{2t}] \leq |I|B^{2t} \leq nB^{2t}$ as well. Combining these, we obtain $\sum_{i \in I} c_i \mathbb{E}_{\tilde{\mu}(v)}[\langle w_i^* - \mu, v \rangle^{2t}] \leq 2^{2t-1} \cdot 17nB^{2t}$. So, $w_i - \mu$ has small $2t$ th pseudomoment.

In addition, since (5.7) is small we know that $\sum_{i \in I} c_i \tilde{\mathbb{E}}_{\zeta_i(z_i)}[\langle w_i^*, z_i \rangle^{2t}] \leq 16r^{2t}/\alpha$ whenever the z_i have small $2t$ th pseudomoment. Putting these together, we get

$$\sum_{i \in I} \|w_i^* - \mu\|_2^{4t} = \sum_{i \in I} \langle w_i^* - \mu, w_i^* - \mu \rangle^{2t} \quad (5.11)$$

$$\leq 2^{2t-1} \sum_{i \in I} \langle w_i^* - \mu, \mu \rangle^{2t} + \langle w_i^* - \mu, w_i^* \rangle^{2t} \quad (5.12)$$

$$\leq 2^{2t-1} \left((2^{2t-1} 17nB^{2t}) \cdot \|\mu\|_2^{2t} + (2^{2t-1} 17nB^{2t}) \cdot (16r^{2t}/\alpha) \right) \quad (5.13)$$

$$\leq 2^{4t-2} \cdot 289n(Br)^{2t}/\alpha. \quad (5.14)$$

In particular, we have

$$\frac{1}{|I|} \sum_{i \in I} \|w_i^* - \mu\|_2^{4t} \leq O(1) \cdot (4Br)^{2t} \cdot (n/\alpha|I|) = O((4Br)^{2t}/\alpha^2). \quad (5.15)$$

This, together with the earlier bound $\frac{1}{|I|} \sum_{i \in I} (1 - c_i) \leq \frac{1 - |I|/n}{3}$, yields (5.8), which establishes the proposition. \square

5.3 Proving Theorem 5.4 via Re-Centering

Proposition 5.7 has a dependence on r , which we would like to get rid of. If $\alpha = 1 - \varepsilon$ where $\varepsilon \leq \frac{1}{4}$, then there is a fairly simple strategy for doing so, based on approximately recovering the mean μ

(with error that depends on \sqrt{r}) and re-running Algorithm 2 after re-centering the points around the approximate mean.

The following key fact about the output of Algorithm 2 will aid us in this (proof in Section D):

Proposition 5.8. *Let $w_{1:n}, c_{1:n}$ be the output of Algorithm 2. Let $S \subseteq \{1, \dots, n\}$ be a subset of at least $\frac{n}{2}$ points such that (i) $c_i \geq \frac{1}{4}$ for all $i \in S$, and (ii) $\|w_i - \hat{\mu}\|_2 \leq \rho$ for all $i \in S$, where $\hat{\mu}$ is the mean of S .*

Then, for any subset S' of S with $|S'| \geq (1 - \varepsilon_1)|S|$ and $\varepsilon_1 \leq \frac{1}{2}$, the mean of S' is close to the mean $\hat{\mu}$ of S :

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \hat{\mu}) \right\|_2 \leq 2\varepsilon_1 \rho + \mathcal{O}(B\varepsilon_1^{1-1/2t}). \quad (5.16)$$

The property that all large subsets of S have mean close to the mean of S is called *resilience*, and was introduced in [SCV18]. There it is shown that one can obtain a good approximation to the true mean μ by finding any large resilient set. Here we repeat the short argument of this fact, and show that a large resilient set can be found efficiently.

First note that when $\varepsilon = 1 - \alpha \leq \frac{1}{2}$, Proposition 5.7 guarantees that a large subset S_0 of at least $1 - \varepsilon$ of the points in I satisfies the conditions of Proposition 5.8 with radius $\rho = C\sqrt{rB\varepsilon^{-1/2t}}$ for some appropriate constant C . Moreover, we can efficiently find such a set (with radius at most 4ρ) as follows:

- Find any point i_0 such that there are at least $(1 - \varepsilon)n$ points j with $\|w_{i_0} - w_j\|_2 \leq 2\rho$ and $c_j \geq \frac{1}{4}$.
- Output these $(1 - \varepsilon)n$ indices j as the set S .

We know that some set S will be found by this procedure (since any $i_0 \in S_0$ will satisfy the conditions). Moreover, by the triangle inequality all points in S will be distance at most 4ρ from the mean of S .

Now, both S and S_0 satisfy the assumptions of Proposition 5.8, and their intersection $S \cap S_0$ has size at least $(1 - 2\varepsilon)n \geq (1 - 2\varepsilon)|S|$. Therefore, assuming $\varepsilon \leq \frac{1}{4}$, the mean μ of S is close to the mean $\hat{\mu}$ of S_0 (since they are both close to the mean of $S \cap S_0$ by Proposition 5.8). This yields:

$$\|\mu - \hat{\mu}\|_2 = \mathcal{O}(\varepsilon\rho + B\varepsilon^{1-1/2t}) = \mathcal{O}(\varepsilon\sqrt{rB\varepsilon^{-1/2t}} + B\varepsilon^{1-1/2t}). \quad (5.17)$$

By finding any set S_0 as above and computing its mean, we obtain a point $\hat{\mu}$ that satisfies (5.17). We can then re-run Algorithm 2 centered at the new point $\hat{\mu}$ (i.e., shifting all of the points by $\hat{\mu}$ as a pre-processing step). When we do this, we can use a smaller radius $r' = C \cdot (\sqrt{rB\varepsilon^{2-1/2t}} + B\varepsilon^{1-1/2t})$ for an appropriate universal constant C . After repeating this a small number of times, we eventually end up with a radius that is $\mathcal{O}(B\varepsilon^{1-1/2t})$. This yields Theorem 5.4.

5.4 Proving Theorem 5.5 via Re-Clustering

We now turn to the more complicated case where potentially $\alpha \ll 1$. Proposition 5.7 says that Algorithm 2 somewhat succeeds at identifying the true mean: while initially we only knew that $\|\mu\|_2 \leq r$, the output of Algorithm 2 has at least $\frac{\alpha n}{2}$ points such that $\|w_i - \mu\|_2 \leq \mathcal{O}(\sqrt{rB\alpha^{-1/m}})$. However, our goal is to obtain a point w such that $\|w - \mu\|_2 = \mathcal{O}(B\alpha^{-1/m})$, with no dependence on r . To obtain this stronger bound, the intuition is to sub-divide the w_i into clusters of radius $\mathcal{O}(\sqrt{rB\alpha^{-1/m}})$, and then recursively apply Algorithm 2 to each cluster.

As in the previous section, we will make use of the resilience property in order to find a good clustering. In this case, we need the following analog of Proposition 5.8 (proof in Section D):

Proposition 5.9. Let w_1, \dots, w_n be the output of Algorithm 2. Let $S \subseteq \{1, \dots, n\}$ be a subset of points such that (i) $c_i \geq \frac{1}{4}$ for all $i \in S$, and (ii) $\|w_i - \hat{\mu}\|_2 \leq \rho$ for all $i \in S$ for some $\hat{\mu}$.

Then, for any subset S' of S with $|S'| \geq \beta n$, the mean of S' is close to $\hat{\mu}$:

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \hat{\mu}) \right\|_2 \leq \rho + \mathcal{O}(B/\beta^{1/2t}). \quad (5.18)$$

Again, sets S satisfying the conclusion of Proposition 5.9 are called resilient sets. Similarly to the case above where $\alpha = 1 - \varepsilon$, [SCV18] shows that, by covering $\{1, \dots, n\}$ with approximately disjoint resilient sets, one can obtain a list of $\mathcal{O}(1/\alpha)$ candidate means $\hat{\mu}$ such that one of the candidates is close to the true mean μ . We will review their argument here, extending it to suit our purposes.

Our final algorithm is going to run Algorithm 2 centered at a number of different points $\tilde{\mu}_1, \dots, \tilde{\mu}_b$ and then consolidate their outputs into a small number of clusters (and then iteratively re-run Algorithm 2 at a new set of points based on the centers of those clusters).

Specifically, suppose that Algorithm 2 has just been run around points $\tilde{\mu}_1, \dots, \tilde{\mu}_b$, and that some unknown $\tilde{\mu}_{a^*}$ (with $1 \leq a^* \leq b$) is close to the target μ : $\|\tilde{\mu}_{a^*} - \mu\|_2 \leq R$. Let $w_{1:n}^{(1)}, \dots, w_{1:n}^{(b)}$ and $c_{1:n}^{(1)}, \dots, c_{1:n}^{(b)}$ denote the outputs of Algorithm 2 on $\tilde{\mu}_1, \dots, \tilde{\mu}_b$. By Proposition 5.7, we have that $c_i^{(a^*)} \geq \frac{1}{4}$ and $\|w_i^{(a^*)} - \mu\|_2 \leq \rho$ for at least $\frac{\alpha n}{2}$ values of i , with $\rho = C\sqrt{RB/\alpha^{1/t}}$.

Now, consider any maximal collection of sets $S_1, \dots, S_m \subseteq [n]$ with the following properties:

- $|S_j| \geq \frac{\alpha}{4}n$ for all j .
- S_j and $S_{j'}$ are disjoint for $j \neq j'$
- For each j , there is some a and $i_0 \in S_j$ such that $\|w_i^a - w_{i_0}^a\|_2 \leq 2\rho$ for all $i \in S_j$, and moreover $c_i^a \geq \frac{1}{4}$ for all $i \in S_j$.

Note that such a collection can be found efficiently via a greedy algorithm: For each i_0 and a , check if there are at least $\frac{\alpha}{4}n$ indices i which do not yet lie in any of the existing S_j , and such that $\|w_i^a - w_{i_0}^a\|_2 \leq 2\rho$ and $c_i^a \geq \frac{1}{4}$. If there are, we can add those indices i as a new set S_j , and otherwise the collection of existing S_j is maximal. Similarly to [SCV18], we then have:

Proposition 5.10. For each S_j , let $\hat{\mu}_j$ be the mean of S_j . Then, at least one of the $\hat{\mu}_j$ satisfy $\|\hat{\mu}_j - \mu\|_2 \leq C'(\sqrt{RB/\alpha^{1/t}} + B/\alpha^{1/t})$ for some universal constant C' .

Proof. Note that the set $S = \{i \mid c_i^{(a^*)} \geq \frac{1}{4} \text{ and } \|w_i^{(a^*)} - \mu\|_2 \leq \rho\}$ satisfies the necessary conditions to be a valid set S_j , and has size at least $\frac{\alpha n}{2}$. Therefore, by the maximality of the collection S_1, \dots, S_m , at least $\frac{\alpha}{4}n$ points in S must lie in $S_1 \cup \dots \cup S_m$ (as otherwise $S \setminus (S_1 \cup \dots \cup S_m)$ could be added to the collection). By Pigeonhole, at least $\frac{\alpha}{4m}n$ points must lie in a specific S_i , and since the S_j are disjoint we have $m \leq \frac{4}{\alpha}$. Therefore, $|S \cap S_i| \geq \frac{\alpha^2}{16}n$ for some i . But by Proposition 5.9 applied to S_j , the mean of $S \cap S_j$ must be within $\mathcal{O}(\rho + B/\alpha^{1/t})$ of the mean $\hat{\mu}_j$ of S_j . Similarly, by Proposition 5.9 applied to S , it is within $\mathcal{O}(\rho + B/\alpha^{1/t})$ of μ . Therefore, $\|\mu - \hat{\mu}_j\|_2 \leq \mathcal{O}(\rho + B/\alpha^{1/t}) = \mathcal{O}(\sqrt{RB/\alpha^{1/t}} + B/\alpha^{1/t})$, as was to be shown. \square

This leads to Algorithm 3, which iteratively runs Algorithm 2, using Proposition 5.10 to find a set of at most $\frac{4}{\alpha}$ means such that one of them grows increasingly close to μ with each iteration. By repeatedly applying Proposition 5.10, we can show that the output $U = \{\hat{\mu}_1, \dots, \hat{\mu}_m\}$ of Algorithm 3 satisfies $m \leq \frac{4}{\alpha}$ and $\|\hat{\mu}_j - \mu\|_2 \leq \mathcal{O}(B/\alpha^{1/t})$, which yields Theorem 5.5.

Algorithm 3 Algorithm for re-clustering the w_i

```
1: Input:  $x_1, \dots, x_n, \alpha, B$ , and upper bound  $r$  on  $\|\mu\|_2$ 
2: Initialize  $R \leftarrow r, U \leftarrow \{0\}$ .
3:  $r_{\text{final}} \leftarrow \Theta(B\alpha^{-1/m})$ .
4: while  $R \geq r_{\text{final}}$  do
5:    $b \leftarrow 1$ 
6:   for  $\hat{\mu} \in U$  do
7:     Let  $w_{1:n}^{(b)}, c_{1:n}^{(b)}$  be the output of Algorithm 2 centered at  $\hat{\mu}$  with parameters  $B, \alpha, R$ .
8:      $b \leftarrow b + 1$ 
9:   end for
10:  Let  $S_1, \dots, S_m$  be the maximal covering derived from  $w_{1:n}^{(1:b)}, c_{1:n}^{(1:b)}$  as in Proposition 5.10.
11:   $U \leftarrow \{\hat{\mu}_j\}_{j=1}^m$ , where  $\hat{\mu}_j$  is the mean of  $S_j$ 
12:   $R \leftarrow C'(\sqrt{RB/\alpha^{1/t}} + B/\alpha^{1/t})$ .
13: end while
14: return  $U$ 
```

5.5 Theorem 5.1: Tighter bounds for well-separated clusters

Suppose that rather than having a single good set I with associated mean μ , we have k good sets I_1, \dots, I_k , with corresponding means μ_1, \dots, μ_k and sizes $\alpha_1 n, \dots, \alpha_k n$, together with a fraction $\varepsilon = 1 - (\alpha_1 + \dots + \alpha_k)$ of arbitrary outliers. In this case, assuming that the μ_j are *well-separated*, we can get substantially stronger bounds by showing that every S_j obtained in Algorithm 3 is almost entirely a subset of one of the I_j . This is what yields Theorem 5.1.

Proof of Theorem 5.1. We first show that in the final iteration of Algorithm 3, the S_j from Proposition 5.10 must all have their mean $\hat{\mu}_j$ be close to a true mean $\mu_{j'}$. Indeed, a given S_j has at most $\varepsilon n \leq \frac{1}{8}\alpha n$ outlier points, and hence at least $\frac{1}{8}\alpha n$ points from $I_1 \cup \dots \cup I_k$. By Pigeonhole it therefore has at least $\frac{1}{8}\alpha \cdot \alpha_{j'} n$ points from some $I_{j'}$. By Proposition 5.9, the mean of these points is within distance $\mathcal{O}(B/\alpha^{1/2t})$ of $\mu_{j'}$, and within distance $\mathcal{O}(B/\alpha^{1/t})$ of $\hat{\mu}_j$, and so $\|\mu_{j'} - \hat{\mu}_j\|_2 \leq \mathcal{O}(B/\alpha^{1/t})$.

Moreover, each cluster S_j is mostly “pure” in the sense that most points come from a single $I_{j'}$. Indeed, suppose for the sake of contradiction that more than $\frac{1}{4}\delta\alpha\alpha_{j''}$ points came from any $I_{j''} \neq I_{j'}$, with $\delta = \Theta(C_{\text{sep}}^{-2t})$. Then the mean of $S_j \cap I_{j''}$ would be within $\mathcal{O}(B/(\alpha^2\delta)^{1/2t}) = \frac{1}{3}C_{\text{sep}}B/\alpha^{1/t}$ of $\mu_{j''}$, while we already had that the mean of $S_j \cap I_{j'}$ was within $\mathcal{O}(B/\alpha^{1/2t}) \leq \frac{1}{3}C_0B/\alpha^{1/t}$ of $\mu_{j'}$ (for an appropriate constant C_0). But $\|\mu_{j'} - \mu_{j''}\|_2 \geq C_{\text{sep}}B/\alpha^{1/t}$ and S_j has radius $\mathcal{O}(B/\alpha^{1/t}) < \frac{1}{3}C_0B/\alpha^{1/t}$, which yields a contradiction whenever $C_{\text{sep}} \geq C_0$. Summing over j'' , we have that at most $\frac{1}{4}\delta\alpha n \leq \delta|S_j|$ points come from any $I_{j''}$ with $j'' \neq j'$ (here we use $|S_j| \geq \frac{\alpha}{4}n$).

Now, because all of the sets S_j have mean close to some $\mu_{j'}$, and the $\mu_{j'}$ are all far apart, we can consolidate the sets S_j into k new sets $\tilde{S}_1, \dots, \tilde{S}_k$ (by e.g. merging together all S_j whose means are within distance $(C_{\text{sep}}/4) \cdot B/\alpha^{1/t}$). Each \tilde{S}_j will have at most $\delta|\tilde{S}_j|$ points from $\cup_{j' \neq j} I_{j'}$, and will have at most $\varepsilon n = (\varepsilon/\alpha_j)|\tilde{S}_j|$ outliers. We then invoke Theorem 5.4 with a $\varepsilon/\alpha_j + \delta \leq \varepsilon/\alpha + \mathcal{O}(C_{\text{sep}}^{-2t})$ fraction of outliers; as long as $\varepsilon \leq \frac{\alpha}{8}$, the fraction of outliers will be below $\frac{1}{4}$ for sufficiently large C_{sep} , and hence we obtain the claimed error bound. \square

References

- [ABG⁺14] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. R. Voss, *The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures*, Conference on Learning Theory (COLT), 2014. [1](#)
- [AGH⁺13] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, *Tensor decompositions for learning latent variable models*, arXiv (2013). [1](#)
- [AM05] D. Achlioptas and F. McSherry, *On spectral learning of mixtures of distributions*, Conference on Learning Theory (COLT), 2005. [1](#)
- [AS12] P. Awasthi and O. Sheffet, *Improved spectral-norm bounds for clustering*, Approximation, Randomization, and Combinatorial Optimization (2012), 37–49. [1](#)
- [AW15] R. Adamczak and P. Wolff, *Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order*, Probability Theory and Related Fields **162** (2015), 531–586. [2](#), [6](#), [7](#), [9](#), [13](#), [18](#)
- [BBH⁺12a] B. Barak, F. Brandão, A. Harrow, J. Kelner, D. Steurer, and Y. Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications*, Symposium on Theory of Computing (STOC), 2012, pp. 307–326. [6](#)
- [BBH⁺12b] Boaz Barak, Fernando G. S. L. Brandão, Aram W. Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications [extended abstract]*, STOC’12—Proceedings of the 2012 ACM Symposium on Theory of Computing, ACM, New York, 2012, pp. 307–326. MR 2961513 [2](#)
- [BBV08] M. Balcan, A. Blum, and S. Vempala, *A discriminative framework for clustering via similarity functions*, Symposium on Theory of Computing (STOC), 2008. [5](#), [20](#)
- [BCM⁺14] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, *Smoothed analysis of tensor decompositions*, Symposium on Theory of Computing (STOC), 2014. [1](#)
- [BCV14] A. Bhaskara, M. Charikar, and A. Vijayaraghavan, *Uniqueness of tensor decompositions with applications to polynomial identifiability*, Conference on Learning Theory (COLT), 2014. [1](#)
- [BÉ85] D. Bakry and M. Émery, *Diffusions hypercontractives*, Séminaire de Probabilités XIX 1983/84, 1985, pp. 177–206. [12](#)
- [BGMZ18] J. Bardet, N. Gozlan, F. Malrieu, and P. Zitt, *Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence*, Bernoulli **24** (2018), 333–353. [12](#)
- [BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer, *Dictionary learning and tensor decomposition via the sum-of-squares method [extended abstract]*, STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 143–151. MR 3388192 [2](#)

- [BL97] S. Bobkov and M. Ledoux, *Poincaré’s inequalities and Talagrand’s concentration phenomenon for the exponential distribution*, *Probability Theory and Related Fields* **107** (1997), no. 3, 383–400. [13](#)
- [BS10] M. Belkin and K. Sinha, *Polynomial learning of distribution families*, *Foundations of Computer Science (FOCS)*, 2010. [1](#)
- [CSV17] M. Charikar, J. Steinhardt, and G. Valiant, *Learning from untrusted data*, *Symposium on Theory of Computing (STOC)*, 2017. [3](#), [5](#), [7](#), [8](#), [20](#), [22](#)
- [DKK⁺16] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, *Robust estimators in high dimensions without the computational intractability*, *Foundations of Computer Science (FOCS)*, 2016. [1](#), [5](#)
- [DKK⁺17] ———, *Being robust (in high dimensions) can be practical*, *arXiv* (2017). [1](#), [5](#)
- [GHK15] R. Ge, Q. Huang, and S. M. Kakade, *Learning mixtures of Gaussians in high dimensions*, *Symposium on Theory of Computing (STOC)*, 2015. [1](#)
- [GM83] M. Gromov and V. D. Milman, *A topological application of the isoperimetric inequality*, *American Journal of Mathematics* **105** (1983), no. 4, 843–854. [13](#)
- [GVX14] N. Goyal, S. Vempala, and Y. Xiao, *Fourier PCA and robust tensor decomposition*, *Symposium on Theory of Computing (STOC)*, 2014. [1](#)
- [HK13] D. Hsu and S. M. Kakade, *Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions*, *Innovations in Theoretical Computer Science (ITCS)*, 2013. [1](#)
- [HKZ09] D. Hsu, S. M. Kakade, and T. Zhang, *A spectral algorithm for learning hidden Markov models*, *Conference on Learning Theory (COLT)*, 2009. [1](#)
- [KK10] A. Kumar and R. Kannan, *Clustering with spectral norm and the k-means algorithm*, *Foundations of Computer Science (FOCS)*, 2010, pp. 299–308. [1](#)
- [KMV10] A. T. Kalai, A. Moitra, and G. Valiant, *Efficiently learning mixtures of two Gaussians*, *Symposium on Theory of Computing (STOC)*, 2010, pp. 553–562. [1](#)
- [KS17] P. Kothari and D. Steurer, *Outlier-robust moment-estimation via sum-of-squares*, *arXiv* (2017). [3](#)
- [Las01] Jean B. Lasserre, *New positive semidefinite relaxations for nonconvex quadratic programs*, *Advances in convex analysis and global optimization (Pythagorion, 2000)*, *Nonconvex Optim. Appl.*, vol. 54, Kluwer Acad. Publ., Dordrecht, 2001, pp. 319–331. [MR 1846160](#) [10](#)
- [Lat06] R. Latała, *Estimates of moments and tails of Gaussian chaoses*, *The Annals of Probability* **34** (2006), no. 6, 2315–2331. [7](#), [13](#)
- [LRV16] K. A. Lai, A. B. Rao, and S. Vempala, *Agnostic estimation of mean and covariance*, *Foundations of Computer Science (FOCS)*, 2016. [1](#), [5](#)

- [MJC⁺14] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp, *Matrix concentration inequalities via the method of exchangeable pairs*, *The Annals of Probability* **42** (2014), no. 3, 906–945. [6](#), [18](#), [19](#)
- [MNV09] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, *The planar k -means problem is NP-hard*, *International Workshop on Algorithms and Computation* (2009), 274–285. [1](#)
- [MV10] A. Moitra and G. Valiant, *Settling the polynomial learnability of mixtures of Gaussians*, *Foundations of Computer Science (FOCS)*, 2010. [1](#)
- [Nes00] Yurii Nesterov, *Squared functional systems and optimization problems*, *High performance optimization*, *Appl. Optim.*, vol. 33, Kluwer Acad. Publ., Dordrecht, 2000, pp. 405–440. MR 1748764 [10](#)
- [Par00] Pablo A Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, Citeseer, 2000. [10](#)
- [RV17] O. Regev and A. Vijayaraghavan, *On learning mixtures of well-separated gaussians*, *Foundations of Computer Science (FOCS)*, 2017. [1](#), [3](#), [4](#), [6](#)
- [SCV18] J. Steinhardt, M. Charikar, and G. Valiant, *Resilience: A criterion for learning in the presence of arbitrary outliers*, *Innovations in Theoretical Computer Science (ITCS)*, 2018. [1](#), [3](#), [5](#), [7](#), [8](#), [9](#), [22](#), [24](#), [25](#)
- [Sho87] N. Z. Shor, *Quadratic optimization problems*, *Izv. Akad. Nauk SSSR Tekhn. Kibernet.* (1987), no. 1, 128–139, 222. MR 939596 [10](#)
- [Tro15] J. A. Tropp, *An introduction to matrix concentration inequalities*, *Foundations and Trends in Machine Learning* **8** (2015), 1–230. [18](#)
- [VW02] S. Vempala and G. Wang, *A spectral algorithm for learning mixture models*, *Foundations of Computer Science (FOCS)*, 2002. [1](#)

A Proof of Lemma 4.7

It suffices to prove the following four inequalities:

$$u \langle M_a \otimes M_b, v^{\otimes(a+b)} \rangle \leq_{a+b} \langle M_{a-1} \otimes M_{b+1}, v^{\otimes(a+b)} \rangle \quad \text{if } a, b \text{ are odd, } a \leq b, \text{ and } u \in \{\pm 1\}. \quad (\text{A.1})$$

$$u \langle M_a \otimes M_b, v^{\otimes(a+b)} \rangle \leq_{a+b} \langle M_{a-2} \otimes M_{b+2}, v^{\otimes(a+b)} \rangle \quad \text{if } a, b \text{ are even, } a \leq b, \text{ and } u \in \{\pm 1\}. \quad (\text{A.2})$$

$$u \langle M_a \otimes M_b, v^{\otimes(a+b)} \rangle \leq_{a+b} \langle M_{a+b}, v^{\otimes(a+b)} \rangle \quad \text{if } a + b \text{ is even and } u \in \{\pm 1\}. \quad (\text{A.3})$$

$$\begin{aligned} \langle A \otimes B, v^{\otimes(a+b)} \rangle &\leq_{a+b} \langle A' \otimes B', v^{\otimes(a+b)} \rangle && \text{if } 0 \leq_a \langle A, v^{\otimes a} \rangle \leq_a \langle A', v^{\otimes a} \rangle \text{ and} \\ &&& 0 \leq_b \langle B, v^{\otimes b} \rangle \leq_b \langle B', v^{\otimes b} \rangle. \end{aligned} \quad (\text{A.4})$$

Given (A.1-A.4), we can use (A.3) to reduce the number of terms M_i in the tensor product in the expression in the claim until there are at most 2 left (such that each term M_i has $i \leq 2t - 2$). We then use (A.1) and (A.2) to iteratively make the indices in the final two terms smaller and larger respectively, until one is 2 and the other is $2t - 2$. This will give us the claim.

Thus, it suffices to establish the inequalities above.

Proof of (A.4): follows from the factorization $A' \otimes B' - A \otimes B = A' \otimes (B' - B) + (A' - A) \otimes B$.

Proof of (A.1): first we need to show that $x^r - ux^{r-1}y - uxy^{r-1} + y^r \succeq_{\text{sos}} 0$ (as a polynomial in x and y) for any even r and $u \in \{\pm 1\}$. We can WLOG take $u = 1$, as if $u = -1$ we just apply the same argument to the variables $(x, -y)$ instead of (x, y) . For $u = 1$ we have

$$x^r - x^{r-1}y - xy^{r-1} + y^r = (x - y)(x^{r-1} - y^{r-1}) \quad (\text{A.5})$$

$$= (x - y)^2(x^{r-2} + x^{r-1}y + x^{r-2}y^2 + \dots + y^{r-2}) \quad (\text{A.6})$$

$$= \frac{1}{2}(x - y)^2((x^{r/2-1})^2 + \sum_{i=1}^{r/2-2} (x^{r/2-i}y^{i-1} + x^{r/2-i-1}y^i)^2 + (y^{r/2-1})^2) \quad (\text{A.7})$$

$$\succeq_{\text{sos}} 0, \quad (\text{A.8})$$

as claimed. Now (A.1) follows as it is equivalent to $x^{a-1}y^{b+1} - ux^a y^b - uxy^a + x^{b+1}y^{a-1}$, which we get by setting $r = b - a + 2$ and multiplying by $x^{a-1}y^{a-1}$.

Proof of (A.2): note that the $u = -1$ case is trivial as then the left-hand-side is $\leq_{\text{sos}} 0$ while the right-hand-side is $\geq_{\text{sos}} 0$. So we can assume $u = 1$. Similarly to the previous proof, it suffices to show that $x^r - x^{r-2}y^2 - x^2y^{r-2} + y^r \succeq_{\text{sos}} 0$. This is actually easier than before:

$$x^r - x^{r-2}y^2 - x^2y^{r-2} + y^r = (x^2 - y^2)(x^{r-2} - y^{r-2}) \quad (\text{A.9})$$

$$= (x^2 - y^2)^2(x^{r-4} + x^{r-6}y^2 + \dots + y^{r-2}) \succeq_{\text{sos}} 0. \quad (\text{A.10})$$

Again, setting $r = b - a + 4$ and multiplying by $x^{a-2}y^{a-2}$ yields the desired result.

Proof of (A.3): follows from repeated application of (A.1) and (A.2). \square

B Proof of Theorem 3.4

First, we establish the product property. Suppose that X_1, X_2 both satisfy (3.1) with constants σ_1, σ_2 which are both at most σ . Supposing that $X_1 \in \mathbb{R}^{d_1}$ and $X_2 \in \mathbb{R}^{d_2}$, and that $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$, we have

$$\text{Var}_{x_1, x_2}[f(x_1, x_2)] \stackrel{(i)}{=} \text{Var}_{x_1}[\mathbb{E}_{x_2}[f(x_1, x_2) \mid x_1]] + \mathbb{E}_{x_1}[\text{Var}_{x_2}[f(x_1, x_2) \mid x_1]] \quad (\text{B.1})$$

$$\stackrel{(ii)}{\leq} \sigma^2 \left(\mathbb{E}_{x_1}[\|\nabla_{x_1} \mathbb{E}_{x_2}[f(x_1, x_2)]\|_2^2] + \mathbb{E}_{x_1}[\mathbb{E}_{x_2}[\|\nabla_{x_2} f(x_1, x_2)\|_2^2]] \right) \quad (\text{B.2})$$

$$= \sigma^2 \left(\mathbb{E}_{x_1}[\|\mathbb{E}_{x_2}[\nabla_{x_1} f(x_1, x_2)]\|_2^2] + \mathbb{E}_{x_1, x_2}[\|\nabla_{x_2} f(x_1, x_2)\|_2^2] \right) \quad (\text{B.3})$$

$$\stackrel{(iii)}{\leq} \sigma^2 \left(\mathbb{E}_{x_1, x_2}[\|\nabla_{x_1} f(x_1, x_2)\|_2^2] + \mathbb{E}_{x_1, x_2}[\|\nabla_{x_2} f(x_1, x_2)\|_2^2] \right) \quad (\text{B.4})$$

$$= \sigma^2 \mathbb{E}_{x_1, x_2}[\|\nabla f(x_1, x_2)\|_2^2]. \quad (\text{B.5})$$

Here (i) is the law of total variation, (ii) applies (3.1) for the individual random variables X_1 and X_2 , and (iii) is Jensen's inequality for the Euclidean norm.

Next, we establish the Lipschitz composition property. Suppose that $X \in \mathbb{R}^d$ satisfies (3.1) with constant σ , and that $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ satisfies $\|\Phi(x) - \Phi(x')\|_2 \leq L\|x - x'\|_2$ for all $x, x' \in \mathbb{R}^d$. Then for $f : \mathbb{R}^{d'} \rightarrow \mathbb{R}$, we have

$$\text{Var}_x[f(\Phi(x))] \stackrel{(i)}{\leq} \sigma^2 \mathbb{E}_x[\|\nabla(f \circ \Phi)(x)\|_2^2] \quad (\text{B.6})$$

$$\stackrel{(ii)}{=} \sigma^2 \mathbb{E}_x[\|\frac{\partial \Phi}{\partial x} \nabla f(\Phi(x))\|_2^2] \quad (\text{B.7})$$

$$\leq \sigma^2 \mathbb{E}_x \left[\left\| \frac{\partial \Phi}{\partial x} \right\|_{\text{op}}^2 \|\nabla f(\Phi(x))\|_2^2 \right] \quad (\text{B.8})$$

$$\stackrel{\text{(iii)}}{\leq} \sigma^2 L^2 \mathbb{E}_x \left[\|\nabla f(\Phi(x))\|_2^2 \right]. \quad (\text{B.9})$$

Here (i) is (3.1) applied to the function $f \circ \Phi$, (ii) is the chain rule, and (iii) is the assumed Lipschitz bound on Φ .

Finally, we establish the linear combination property, which follows from the two properties above. Suppose that X_1, X_2 satisfy (3.1) with constants σ_1, σ_2 . Then by the product property, the random variable $(X_1/\sigma_1, X_2/\sigma_2)$ satisfies (3.1) with constant 1. But now the map $(Y_1, Y_2) \mapsto a\sigma_1 Y_1 + b\sigma_2 Y_2$ is Lipschitz with constant $L = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}$, so in particular $aX_1 + bX_2$ satisfies (3.1) with constant $\sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}$. This completes the proof.

C Proof of Lemma 5.6

Note that the optimization problem

$$\min_{w_1, \dots, w_n} \max_{\xi, \zeta_1, \dots, \zeta_n} \sum_{i=1}^n c_i \tilde{\mathbb{E}}_{\xi(v)} [\langle x_i - w_i, v \rangle^{2t}] + \lambda \tilde{\mathbb{E}}_{\zeta_i(z_i)} [\langle w_i, z_i \rangle^{2t}] \quad (\text{C.1})$$

is a convex-concave saddle point problem. In particular, convexity in w follows because the convexity (in v) of the function $\langle x_i - w_i, v \rangle^{2t}$ has a sum-of-squares proof, and similarly for $\langle w_i, z_i \rangle^{2t}$. On the other hand, concavity in ξ, ζ holds because the objective is in fact linear in z_i .

In particular, by the minimax theorem, (C.1) remains unchanged if we swap the order of min and max. However, we also need to check that both directions of the saddle point problem can be efficiently computed.

To do this, we first relax w_1, \dots, w_n to themselves be pseudodistributions $\theta_i(w_i)$. Then the entire objective for (C.1) becomes bilinear in θ and in ξ, ζ . Moreover, since sum-of-squares programs can be represented as semidefinite programs, the relaxed version of (C.1) is in fact a bilinear min-max problem with semidefinite constraints on the two sets of variables. It is a standard fact in optimization that such programs can be solved efficiently to arbitrary precision.

Now, to finish, we note that the relaxation of w_i was not in fact a relaxation at all! Because (C.1) was already convex in w_i , there is always an optimal θ_i that places all of its mass on a single point. In particular, given optimal values $\theta_1^*, \dots, \theta_n^*$, we can simply take $w_i^* = \tilde{\mathbb{E}}_{\theta_i^*(w)}[w]$, which must have at least as good of an objective value for (C.1) (because of the sum-of-squares proof of convexity).

D Proof of Propositions 5.8 and 5.9

The proofs of Proposition 5.8 and 5.9 are almost identical, so we prove them together.

We start with Proposition 5.9. We write

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \hat{\mu}) \right\|_2 = \left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - w_i) + (w_i - \hat{\mu}) \right\|_2 \quad (\text{D.1})$$

$$\leq \rho + \sup_{\|v\|_2 \leq 1} \frac{1}{|S'|} \sum_{i \in S'} \langle x_i - w_i, v \rangle \quad (\text{D.2})$$

$$\leq \rho + \sup_{\|v\|_2 \leq 1} \left(\frac{1}{|S'|} \sum_{i \in S'} \langle x_i - w_i, v \rangle^{2t} \right)^{1/2t} \quad (\text{D.3})$$

$$\leq \rho + \sup_{\|v\|_2 \leq 1} \left(\frac{4}{|S'|} \sum_{i=1}^n c_i \langle x_i - w_i, v \rangle^{2t} \right)^{1/2t} \quad (\text{D.4})$$

$$= \rho + \left(\frac{4}{|S'|} \mathcal{O}(nB^{2t}) \right)^{1/2t} \quad (\text{D.5})$$

$$= \rho + \mathcal{O}(B/\beta^{1/2t}) \quad (\text{D.6})$$

as was to be shown.

We now consider Proposition 5.8. We have

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \hat{\mu}) \right\|_2 = \left\| \frac{1}{|S'|} \sum_{i \in S \setminus S'} (x_i - \hat{\mu}) \right\|_2 \quad (\text{D.7})$$

$$= \frac{\varepsilon_1}{1 - \varepsilon_1} \left\| \frac{1}{|S \setminus S'|} \sum_{i \in S \setminus S'} (x_i - \hat{\mu}) \right\|_2 \quad (\text{D.8})$$

$$\stackrel{(i)}{\leq} \frac{\varepsilon_1}{1 - \varepsilon_1} (\rho + \mathcal{O}(B/(\varepsilon_1 \beta)^{1/2t})) \quad (\text{D.9})$$

$$\leq 2\varepsilon_1 \rho + \mathcal{O}(B \varepsilon_1^{1-1/2t} \beta^{-1/2t}), \quad (\text{D.10})$$

where (i) simply applies Proposition 5.9.